Recommendations for Video Event Recognition using Concept Vocabularies

Amirhossein Habibian, Koen E. A. van de Sande, and Cees G. M. Snoek
ISLA, Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH, Amsterdam, The Netherlands
{a.habibian, ksande, cgmsnoek}@uva.nl

ABSTRACT

Representing videos using vocabularies composed of concept detectors appears promising for event recognition. While many have recently shown the benefits of concept vocabularies for recognition, the important question what concepts to include in the vocabulary is ignored. In this paper, we study how to create an effective vocabulary for arbitrary-event recognition in web video. We consider four research questions related to the number, the type, the specificity and the quality of the detectors in concept vocabularies. A rigorous experimental protocol using a pool of 1,346 concept detectors trained on publicly available annotations, a dataset containing 13,274 web videos from the Multimedia Event Detection benchmark, 25 event groundtruth definitions, and a state-of-the-art event recognition pipeline allow us to analyze the performance of various concept vocabulary definitions. From the analysis we arrive at the recommendation that for effective event recognition the concept vocabulary should *i)* contain more than 200 concepts, *ii)* be diverse by covering *object*, *action*, *scene*, *people*, *animal* and *attribute* concepts, *iii)* include both general and specific concepts, and *iv)* increase the number of concepts rather than improve the quality of the individual detectors. We consider the recommendations for video event recognition using concept vocabularies the most important contribution of the paper, as they provide guidelines for future work.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Event recognition, Concept representation

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.

1. INTRODUCTION

We consider the problem of event recognition in arbitrary web video. Among the many challenges involved, resulting from the uncontrolled recording condition of web videos and the large variations in the visual appearances of events, probably one of the most fundamental questions in event recognition is what defines an event in video? The Oxford English dictionary defines an event as "anything that happens". With such a broad definition it is not surprising that the topic has been addressed in the multimedia retrieval community by many researchers from diverse angles [2, 26, 19, 3, 11].

In this paper, we study event representations that contribute to defining events for automatic recognition. We are inspired by findings from cognition, where research has repeatedly shown that humans remember events by their actors, actions, objects, and locations [20]. Studying event representation based on such high-level concepts is now within reach because of the continued progress in supervised concept detection [22] and the availability of labeled training collections like the ones developed in benchmarks like TRECVID [21], ImageNet [5] and several other venues [13, 7]. In this paper, we name the set of available concept detectors as the *vocabulary* and we study its ideal composition for effective recognition of events in arbitrary web video.

1.1 Representing Events in Video

The state-of-the-art in event recognition represents a video in terms of low-level audiovisual features [15, 23, 14, 9]. In general, these methods first extract from the video various types of static and/or dynamic features, *e.g.*, color SIFT variations [25], MFCC [9], and Dense Trajectories [15]. Second, the descriptors are quantized and aggregated [15]. The robustness and efficiency of various low-level features for recognizing events are evaluated in [23, 9]. Despite their good recognition performance, especially when combined together [14, 9, 15], low-level features are incapable of providing understanding of the semantic structure present in an event. Hence, it is not easy to derive how these event definitions arrive at their recognition. Therefore, essentially different representations are needed for events. We focus on high-level representations for event recognition.

Inspired by the previous works in object recognition [24, 10], scene recognition [10, 17] and activity recognition [18], others have also explored high-level representations for recognition of events [12, 1, 27, 8]. In all these works, the video is represented as the output of numerous pre-trained concept detector scores. In [12], for example, Merler *et al.*

Table 1: Examples of videos and human-added textual descriptions, from which we study how humans describe events.

A woman folds and packages a scarf she has made.

A woman points out bones on a skeleton for lab practical for an anatomy class.

A little boy helps a woman put up cobweb Halloween decorations at the front of a house.

People competing in a sand sculpting competition and children playing on the beach.

A mother at a fountain tries to get her daughter to step on the water jets.

arrive at a robust high-level representation of events using 280 concept detectors, which outperforms a low-level audiovisual representations using a state-of-the-art recognition pipeline following three consecutive steps. First, frame extraction, where the videos are decoded and a subset of frames are extracted. Second, concept detection, where a set of pre-trained concept detectors are applied on the extracted frames. Each frame is then represented as a concept vector obtained by concatenating all the detector outputs. Finally, video pooling, where the frame representations are averaged and aggregated into the video level representation. However, in the paper by Merler *et al.*, as well as all the others [1, 27, 8], the question what concepts to include in the vocabulary to represent events is ignored. In this paper, we adopt the event recognition pipeline of [12], but we place special emphasis on what concepts to insert in the vocabulary for effective event recognition.

1.2 What Concepts?

Our study is inspired by the pioneering work of Hauptmann *et al.* [6] who focus on concept vocabularies for broadcast news video. They examined how big the concept vocabulary should be and what concepts should be part of the vocabulary for effective shot retrieval. In their work, the presence and absence of 320 human-annotated concepts was used as the main source for the investigations. By inserting the same amount of noise into each of the human annotations they were able to study news video retrieval accuracy under varying levels of concept quality, arriving at the prediction that 5,000 detectors with modest quality would be sufficient for general-purpose broadcast news video retrieval. However, it is not clear whether their conclusion generalizes to *event* recognition on the challenging domain of unconstrained web video. Regarding the important question what concepts to include in the vocabulary, Hauptmann *et al.* [6] conclude that frequent concepts contribute more to overall news video retrieval performance than rare concepts, but they do not make a distinction with respect to concept type.

In this paper, we start from the analysis by Hauptmann *et al.* [6]. We adopt three of their research questions, as well as their idea to insert (additional) noise into the concepts. However, our work is different with respect to the following five aspects. First, we focus exclusively on events, whereas [6] also considers news use cases like *Find shots of U.S. Maps depicting the electoral vote distribution (blue vs. red state)* and *Find shots of Refugee Camps with women and children visible*. Second, our domain of study is unconstrained web video, rather than the highly structured broadcast television domain. Third, we place special emphasis on the importance of various concept types in the vocabulary (e.g., objects, scenes, actions etc.), rather than considering all concepts equally important. Fourth, in our analysis we do not rely on human concept annotations directly, but instead we use real detector predictions with varying levels of accuracy per concept. Finally, we evaluate retrieval accuracy on video-level rather than shot-level. Because of all these differences, we feel a new study on concept vocabularies is justified.

1.3 Research Questions

Our study on the effectiveness of concept vocabularies for video event recognition, is directed by the following four research questions:

RQ1 *How many concepts to include in the vocabulary?*

RQ2 *What concept types to include in the vocabulary?*

RQ3 *Which concepts to include in the vocabulary?*

RQ4 *How accurate should the concept detectors be?*

As humans remember events by the high level concepts they contain, *viz.,* actors, actions, objects, and locations [20], studying the characteristics of the concepts that humans use to describe events could be inspirational for automated event recognition. Therefore, before describing our experimental protocol to address the research questions, we first study the vocabulary that human uses to describe events in videos.

2. HUMAN EVENT DESCRIPTION

To analyze the vocabulary that humans use to describe events, we utilize a set of textual descriptions written by humans to describe web videos containing events. We process textual descriptions for 13,265 videos, as provided by the TRECVID 2012 Multimedia Event Detection task corpus [21]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by highlighting its dominant concepts. Table 1 illustrates some videos and their corresponding textual descriptions.

After removing stop words and stemming, we end up with 5,433 distinct terms from the 13,265 descriptions making up a human vocabulary for describing events. Naturally, the frequency of these terms varies, as also observed by [6]. Most of the terms seldom occur in event descriptions. Whereas, only a few terms have high term frequencies. To be precise, 50% of the terms occur once in the descriptions and only 2% occurs more than five times. Terms like `man` , `girl`, `perform` and `street` appear most frequent, while `bluefish`, `conductor`, `Mississippi` and `Bulgarian` are instances of less frequent terms.

Looking into the vocabulary, we observe that the terms used in human description can be mapped to five distinct

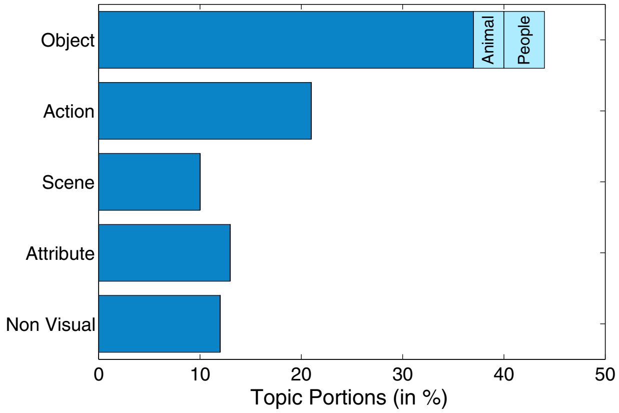

Figure 1: We divide the human vocabulary for describing events into concept types containing objects, actions, scenes, attributes and non-visual concepts. Our analysis reveals that objects and actions constitute 65% of the human vocabulary when describing events.

concept types as typically used in the multimedia and computer vision literature: *objects*, *actions*, *scenes*, *visual attributes* and *non visual concepts*. We manually assign each vocabulary term into one of these five types. After this exercise we observe that 44% of the terms refer to *objects*. Moreover, we note that a considerable number of objects are dedicated to various types of *animals* and *people*; *i.e.*, `lion`, and `teen`. About 21% of the terms depict *actions*, like `walking`. Approximately 10% of the concept types are about *scenes*, such as `kitchen`. *Visual attributes* cover about 13% of the terms; *i.e.*, `white`, `flat`, and `dirty`. The remaining 12% of the terms belong to concepts, which are *not visually depictable*; *i.e.*, `poem`, `problem`, and `language`. We summarize the statistics of our human event descriptions in Figure 1.

We observe that when describing video events, humans use terms with varying generalizations. Some terms are very specialized so refer to specific objects; like, `salmon`, `cheesecake` and `sand castle`. While other terms are more general, so refer to broader set of concepts; like `human`, `vegetation` and `outdoor`. We analyze the generalization of the vocabulary terms using their depth in the WordNet hierarchy. In this hierarchy, the terms are structured based on their hypernym/hyponym relations, so the more specialized terms are placed at the deeper levels. Our study shows that the 5,433 vocabulary terms have an average depth of 9.07±5.29. The high variance in term depths indicates that the human vocabulary to describe events is composed of both specific and general terms.

To summarize, we observe that the vocabulary that humans use to describe events is composed of a few thousand words, derived from five distinct concept types: *objects*, *actions*, *scenes*, *visual attributes* and *non visual concepts*. Moreover, we observe that the vocabulary contains both specific and general concepts. Strengthened by these observations about the human vocabulary for describing events, we design four experiments to answer our research questions on the ideal composition for recognizing events in arbitrary web video.

Table 2: Number of videos in the dataset used for our experiments split per event. Partitioning available from *http://www.mediamill.nl/datasets*.

	Train Set		**Test Set**	
Event	Positives	Negatives	Positives	Negatives
Attempting board trick	98	8742	49	4376
Feeding animal	75	8745	48	4377
Landing fish	71	8769	36	4389
Wedding ceremony	69	8771	35	4390
Working wood working project	79	8761	40	4385
Birthday party	121	8719	61	4364
Changing vehicle tire	75	8765	37	4388
Flash mob gathering	115	8725	58	4367
Getting vehicle unstuck	85	8755	43	4382
Grooming animal	91	8749	46	4379
Making sandwich	83	8757	42	4383
Parade	105	8735	50	4375
Parkour	75	8765	38	4387
Repairing appliance	85	8755	43	4382
Working sewing project	86	8754	43	4382
Attempting bike trick	43	8797	22	4403
Cleaning appliance	43	8797	22	4403
Dog show	43	8797	22	4403
Giving directions location	43	8797	22	4403
Marriage proposal	43	8797	22	4403
Renovating home	43	8797	22	4403
Rock climbing	43	8797	22	4403
Town hall meeting	43	8797	22	4403
Winning race without vehicle	43	8797	22	4403
Working metal crafts project	43	8797	22	4403

3. EXPERIMENTAL SETUP

To answer the research questions raised in the introduction of the paper, we create a rigorous empirical setting. First, we introduce the video dataset used to evaluate the event recognition experiments. Then we explain the pool of concept detectors, which we employ to create vocabularies. Finally, the pipeline used for event recognition using concept vocabularies is presented.

3.1 Video Dataset

For the event recognition experiments, we rely on the video corpus from the TRECVID 2012 Multimedia Event Detection task [21]. To the best of our knowledge this is the largest publicly available video corpus in the literature for event recognition. The corpus consists of over 1,500 hours of user-generated video with a large variation in quality, length and content. Moreover, it comes with ground truth annotations at video level for 25 real-world events, including life events, instructional events, sport events, etc.. We extract two partitions consisting of 8,840 and 4,434 videos from the development set of the corpus. Development set is the annotated part of the corpus, which is suitable to develop and validate the methods. In this paper we use the larger partition as the training set, on which we train our event recognizers, and we report all results on the smaller partition. We summarize the training and test set statistics of the video dataset per event in Table 2.

3.2 Implementation Details

Concept Vocabulary To create the vocabularies, we need a comprehensive pool of concept detectors. We build this pool of detectors using the human annotated training data from two publicly available resources: the TRECVID 2012 Semantic Indexing task [21] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [4]. The former has annotations for 346 semantic concepts on 400,000 keyframes from web videos. The latter has annotations for 1,000 semantic concepts on 1,300,000 photos. The categories are

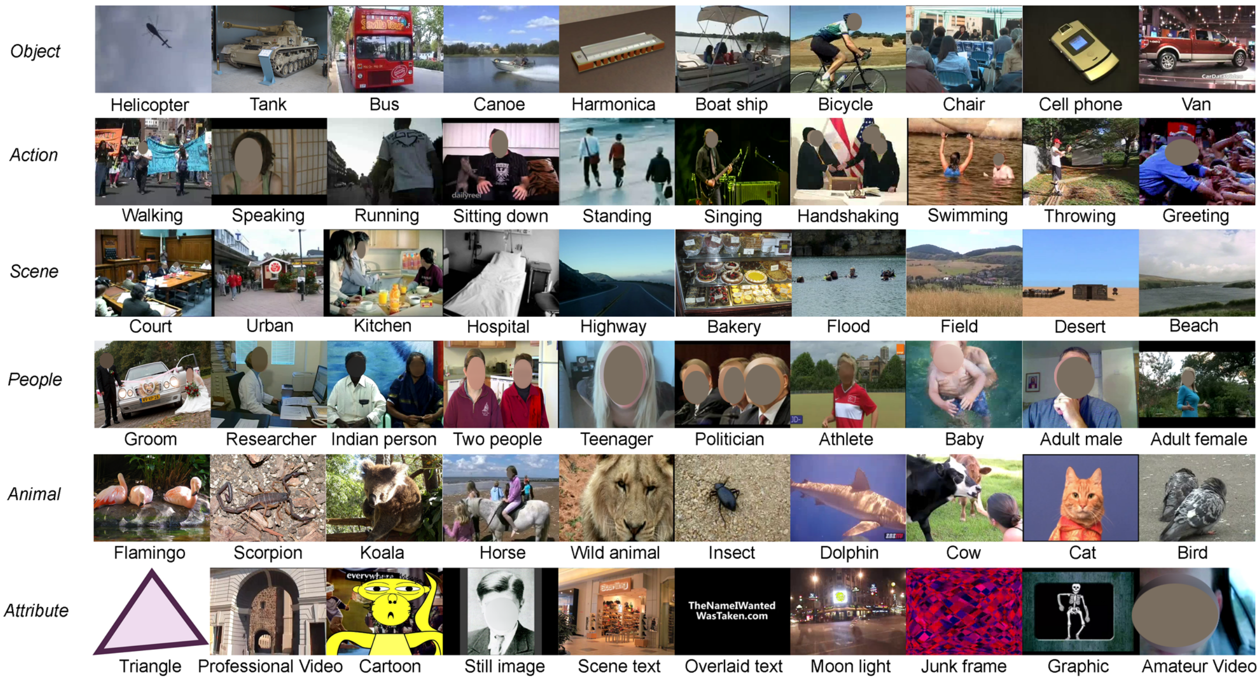

Figure 2: Random examples of the 1,346 concept detectors included in the overall vocabulary used in our experiments, grouped by the concept type.

quite diverse and include concepts from various types; *i.e.*, *object*, *scene* and *action*.

Leveraging the annotated data available in these datasets, we train 1,346 concept detectors in total. We follow the state-of-the-art for our implementation of the concept detectors. We use densely sampled SIFT, OpponentSIFT and C-SIFT descriptors [25] with Fisher vector coding [16]. The visual vocabulary used has a size of 256 words. As a spatial pyramid we use the full image and three horizontal bars. The feature vectors representing the training images form the input for a fast linear Support Vector Machine.

As summarized in Figure 1, the concepts that humans use to describe events are derived from *object*, *action*, *scene*, *attributes* and *non visual* concept types. The *non visual* concepts cannot be detected by their visual features, so we exclude them from our study. Regarding to the importance of the actors in depicting events [20], as well as their high frequency in human descriptions, we consider *people* and *animal* as extra concept types in our experiments. Inspired by this composition, we divide our concept pool by manually assigning them to one of these types. Consequently, we end up with the following concept types: *object* containing 706 concepts, *action* containing 36 concepts, *scene* containing 135 concepts, *people* containing 83 concepts, *animal* containing 338 concepts and *attribute* containing 48 concepts. Figure 2 gives an overview of the concept types and example instances.

Event Recognition In the event recognition experiments, we follow the pipeline proposed in [12]. We decode the videos by uniformly extracting one frame every two seconds. Then all the concept detectors are applied on the extracted frames. Concatenating the detector outputs, each frame is represented by a concept vector. Finally the frame representations are aggregated into a video level representation by averaging and normalization. On top of this concept vocabulary representation per video, we use again a linear SVM classifier to train the event recognizers.

4. EXPERIMENTS

We perform four experiments to address our research questions. Each concept vocabulary used in the experiments is evaluated based on its performance in recognizing events using the pipeline and evaluation protocol described in section 3. Moreover, the vocabularies are all derived from the concept pool introduced in section 3.2.

- **Experiment 1: How many concepts to include in the vocabulary?** To study this question, we create several vocabularies with varying sizes and evaluate their performance for recognizing events. Each vocabulary is made of a random subset of the concept detectors from the concept pool. To compensate for possible random effects, all experiments are repeated 50 times and the results are averaged.

- **Experiment 2: What concept types to include in the vocabulary?** We look into this question by comparing two types of vocabularies: *(i) single type* vocabularies, where all concepts are derived from one type and *(ii) joint type* vocabularies, where concepts are derived from all available concept types. We perform this experiment for six kinds of single type vocabularies: *object*, *action*, *scene*, *people*, *animal* and *attribute* types respectively.

  To make the single type and joint type vocabularies more comparable, we force the vocabularies to be of equal size. We do so by randomly selecting the same number of concepts from the concept pool. All the experiments are repeated 500 times to balance possible random effects.

- **Experiment 3: Which concepts to include in the vocabulary?** In this experiment, we investigate whether the concept vocabulary for event recognition should be made of general or specific concepts. We manually extract two sets of general and specific concepts from the concept pool. The former contains 149 general concepts, *i.e.*, `vegetation`, `human` and `man made thing`, and the latter contains 619 specific concepts, *i.e.*, `religious figure`, `emergency vehicle` and `pickup truck`. The rest of concepts, which are not clearly general or specific, are not involved in this experiment. Using these sets we compare three types of vocabularies: *(i) general* vocabulary in which all the concepts are general, *(ii) specific* vocabulary in which all the concepts are specific and *(iii) mixture* vocabulary in which the concepts are randomly selected from both general and specific concept sets. We repeated this experiment for different vocabulary sizes and found that the results remained stable. The reported results are obtained for a vocabulary size of 70, averaged over 500 repetitions.

- **Experiment 4: How accurate should the concept detectors be?** In this experiment, we decrease the detector accuracies by introducing noise into the concept prediction scores. We gradually increase the amount of noise and measure how the event recognition performance responds.

  The output of each concept detector, as a SVM classifier, is a real value number which is supposed to be larger than +1 and smaller than -1 for respectively positive and negative samples. But in practice, SVM only assigns these values to the samples which are confidently classified, while other samples are assigned to the unconfident area in between -1 and 1. Looking into the concept detector predictions, we observe that most of them are agglomerated in the unconfident area. The less accurate a concept detector is, the more samples are assigned to the unconfident area. To simulate the detector accuracy changes, we randomly select predictions and shift them towards center of the unconfident area, which has the least decision confidence. We gradually increase the amount of noise and repeat the experiments 50 times to compensate for possible random factors.

Each experiment results in a ranking of the videos from the test set based on the probability that the video contains the event of interest. As the evaluation criterion for these ranked lists, we employ average precision (AP) which is in wide use for evaluating visual retrieval results [21]. We also report the average performance over all events as the mean average precision (MAP).

5. RESULTS

5.1 Experiment 1: How many?

As shown in Figure 3, adding more concept detectors to the vocabulary improves the event recognition performance. The improvement gain is particularly prevalent for small vocabularies. When increasing the vocabulary from 50 to 300, for example, the MAP increases from 0.125 to 0.221. The improvement is less prevalent when more than 1,000 detectors are part of the vocabulary. When increasing the vocabulary from 1,000 to 1,346 the absolute MAP improvement is only 0.012 on average.

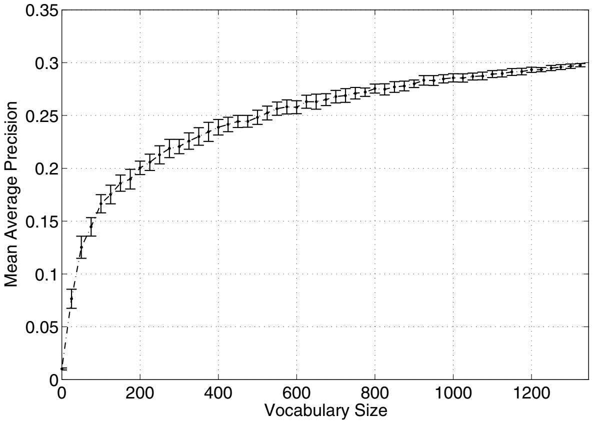

Figure 3: Experiment 1: Increasing the vocabulary size improves the event recognition performance. This improvement is especially prevalent for small vocabularies containing less than 200 concept detectors.

The error bars plotted in Figure 3 indicate the variance in MAPs for various vocabularies. The variance demonstrates that with the same number of concept detectors, some vocabularies perform better than others. In the next two experiments, we study the characteristics of these optimal vocabularies.

Small vocabularies have poor performances in recognizing events. In addition, their efficiency could be rapidly increased by adding few more concepts to them. So, we recommend to include at least 200 concept detectors in the vocabulary.

5.2 Experiment 2: What concept types?

Table 3 compares single type and joint type vocabularies for recognizing events. Comparing the MAPs, we conclude that joint type vocabularies outperform single type vocabularies for all six concept types on average. It demonstrates that when creating the vocabulary, it is better to sample the concept detectors from diverse types. Hence, we need to detect the objects, people, actions and scenes occurring in the video *jointly* to recognize the event properly. In other words, all of the concept types contribute to the recognition of events.

When we analyze individual event recognition results, we observe a few cases exist where a single type vocabulary outperforms the joint type because of the tight connection between the event description and specific concepts. For example, using a single type vocabulary made of *animals* only, we achieve a higher average precision for *"feeding animal"*, *"grooming animal"* and *"dog show"* events in comparison to a joint type vocabulary. Similarly, *"flash mob gathering"*, *"rock climbing"* and *"town hall meeting"* are recognized better by the scene concepts than by the joint vocabulary. Never-

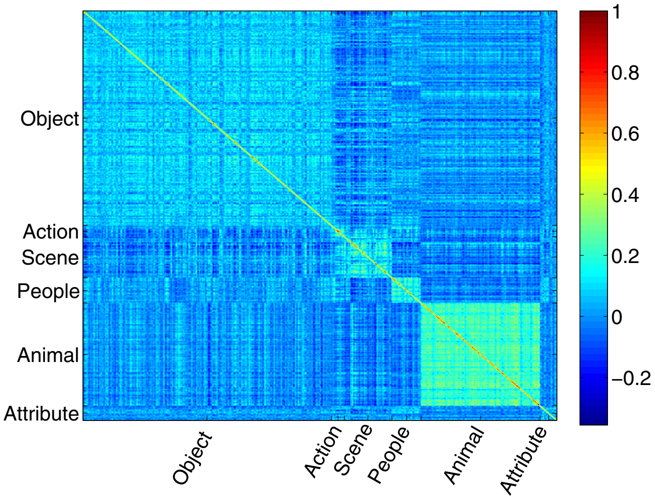

Figure 4: Experiment 2: Correlation between concept detector responses appears especially within a single concept type. Including too many concepts from the same type leads to decreased event recognition performance (matrix smoothed for better viewing).

theless, joint type vocabularies do better than single type vocabularies on average. Therefore, we consider joint type vocabularies more suited for general purpose event recognition.

The performance difference between the single type and joint type vocabularies varies per concept type. For some types, like *animal*, the difference is substantial (0.158 vs. 0.239), while for others, like *action*, it is almost negligible (0.067 vs. 0.076). We attribute the performance difference to at least two reasons. First, our concept detectors are trained on global image level, so they contain considerable amount of contextual information. Consequently, some single types may contain a wide sample of contextual information including 'semantic overlap' from other concept types. The *action*, for example, may contain action detectors in varying scenes using various objects. Second, when creating many concept detectors for a similar type, it is likely the detectors will be correlated to each other, especially for the less diverse types. To clarify this observation we plot the correlation between concept detectors within a concept type in Figure 4. As shown in this figure, the highly correlated concepts tend to belong to the same concept type. Therefore, including too many concepts from the same type in a vocabulary, especially from the less diverse concept types like *animal* and *people*, leads to correlated concepts and should be avoided.

We recommend to make the vocabulary diverse by including concepts from various concept types and to limit the number of concepts for the less diverse types.

5.3 Experiment 3: Which concepts?

Table 4 compares three types of vocabularies: specific, general and mixture. According to the MAPs, the general vocabulary performs better than the specific vocabulary, but the mixture vocabulary is the best overall performer.

Table 4: Experiment 3: Comparison of specific, general and mixture vocabularies. The results demonstrate that the general vocabulary outperforms the specific vocabulary on average. The best results are obtained when the vocabulary consists of both general and specific concepts.

Event	Specific	General	Mixture
Attempting board trick	0.090	0.108	**0.130**
Feeding animal	0.041	0.042	**0.045**
Landing fish	0.113	0.107	**0.139**
Wedding ceremony	0.071	0.140	**0.164**
Working wood working project	**0.083**	0.065	0.073
Birthday party	0.078	0.135	**0.138**
Changing vehicle tire	0.058	0.062	**0.071**
Flash mob gathering	0.301	0.284	**0.337**
Getting vehicle unstuck	0.195	0.246	**0.282**
Grooming animal	0.064	0.079	**0.081**
Making sandwich	0.059	0.089	**0.119**
Parade	0.073	**0.203**	0.161
Parkour	0.104	**0.226**	0.210
Repairing appliance	**0.111**	0.098	0.101
Working sewing project	0.076	0.075	**0.082**
Attempting bike trick	0.044	0.080	**0.090**
Cleaning appliance	**0.125**	0.092	0.123
Dog show	0.219	0.178	**0.230**
Giving directions location	0.028	0.019	**0.053**
Marriage proposal	0.013	0.017	**0.025**
Renovating home	0.023	0.074	**0.083**
Rock climbing	0.178	0.156	**0.194**
Town hall meeting	0.064	**0.226**	0.158
Winning race without vehicle	0.102	0.102	**0.117**
Working metal crafts project	**0.040**	0.021	0.036
Mean	0.094	0.117	**0.130**

We observe that for four events a specific vocabulary outperforms the others: *"working wood working project"*, *"repairing appliance"*, *"cleaning appliance"* and *"working metal crafts project"*. For these events, there are some specific and discriminative concepts available in the vocabulary. For example, `lumber mill`, `crate` and `circular saw` concepts for *"working wood working project"* and `washing machine`, `refrigerator` and `microwave` concepts for *"repairing appliance"*. While the specific concepts may be distinctive for recognizing some events, the concepts typically occur in only few videos. Hence, they are absent in most videos and do not contribute much to event recognition. Therefore, if the vocabulary consists of specific concepts only, it will perform well in recognizing the events relevant to those concepts, but it will perform poor for other events. In contrast to the specific concepts, general concepts occur in a large numbers of videos. Although these concepts are not discriminative individually, taking several of them together into a vocabulary makes the event recognition better than using a specific vocabulary. Since it is able to simultaneously utilize distinctive specific concepts and general concepts, the best performance is obtained when the vocabulary contains a mixture of both specific and general concepts.

We recommend to insert both general and specific concepts into the event recognition vocabulary.

5.4 Experiment 4: How accurate?

As expected, the results in Figure 5 demonstrate event recognition performance degrades by adding more noise to the concept detector predictions in the vocabulary. When the noise amount is rather small, *i.e.*, up to 30%,, the event recognition remains relatively robust. For a vocabulary containing 1,346 concepts, the performance drops by only 3%

Table 3: Experiment 2: Comparison of single type and joint type vocabularies for event recognition. Each column pair compares a single and joint type vocabulary. To make the vocabularies more comparable within a concept type, we force them to be of equal size. Note that the number of concept detectors (in parenthesis) varies per concept type, so comparison across concept types should be avoided. The results demonstrate that for all the six concept types, joint type vocabularies outperform single type vocabularies on average.

	Concept Type											
Event	**Object**(670)		**Action**(34)		**Scene**(128)		**People**(78)		**Animal**(321)		**Attribute**(45)	
	Single	*Joint*	*Single*	*Joint*	*Single*	*Joint*	*Single*	*Joint*	*Single*	*Joint*	*Single*	*Joint*
Attempting board trick	**0.368**	0.348	0.056	**0.073**	0.115	**0.169**	0.065	**0.119**	0.120	**0.271**	**0.082**	0.079
Feeding animal	0.035	**0.044**	0.029	**0.074**	0.024	**0.042**	0.040	**0.041**	**0.073**	0.045	**0.055**	0.037
Landing fish	0.337	**0.423**	0.055	**0.076**	0.157	**0.246**	0.074	**0.182**	0.323	**0.360**	0.054	**0.111**
Wedding ceremony	0.493	**0.520**	0.054	**0.073**	0.139	**0.193**	**0.141**	0.119	0.162	**0.388**	0.040	**0.070**
Working wood working project	0.194	**0.203**	0.029	**0.040**	0.074	**0.101**	**0.118**	0.072	0.116	**0.167**	0.032	**0.048**
Birthday party	0.264	**0.277**	0.098	**0.099**	0.115	**0.174**	**0.138**	0.131	0.139	**0.239**	0.058	**0.095**
Changing vehicle tire	0.171	**0.174**	0.034	**0.054**	0.073	**0.105**	0.036	**0.076**	0.054	**0.153**	0.043	**0.052**
Flash mob gathering	0.471	**0.494**	**0.257**	0.212	**0.349**	0.304	0.321	**0.337**	0.415	**0.475**	**0.273**	0.251
Getting vehicle unstuck	0.330	**0.362**	0.092	**0.138**	0.186	**0.268**	0.110	**0.217**	0.294	**0.338**	0.069	**0.154**
Grooming animal	0.126	**0.149**	0.033	**0.070**	0.129	**0.147**	0.075	**0.080**	**0.146**	0.127	**0.075**	0.068
Making sandwich	0.178	**0.197**	0.023	**0.061**	0.116	**0.127**	0.050	**0.098**	0.070	**0.176**	0.029	**0.066**
Parade	0.268	**0.304**	**0.169**	0.119	0.215	**0.219**	0.119	**0.182**	0.126	**0.275**	0.093	**0.141**
Parkour	0.398	**0.432**	0.023	**0.063**	0.150	**0.234**	0.034	**0.147**	0.089	**0.356**	0.031	**0.074**
Repairing appliance	0.244	**0.323**	0.063	**0.078**	0.192	**0.224**	0.086	**0.126**	0.104	**0.259**	**0.100**	0.083
Working sewing project	**0.295**	0.252	0.048	**0.075**	0.129	**0.163**	0.107	**0.123**	0.194	**0.238**	0.021	**0.082**
Attempting bike trick	0.480	**0.502**	**0.264**	0.076	**0.250**	0.245	0.037	**0.171**	0.129	**0.392**	0.031	**0.096**
Cleaning appliance	**0.079**	0.064	0.019	**0.039**	0.022	**0.049**	0.021	**0.045**	0.029	**0.058**	0.015	**0.035**
Dog show	0.500	**0.534**	0.093	**0.102**	0.423	**0.455**	0.114	**0.236**	**0.555**	0.512	0.116	**0.122**
Giving directions location	0.029	**0.031**	0.013	**0.027**	0.019	**0.025**	0.011	**0.021**	0.016	**0.029**	0.012	**0.021**
Marriage proposal	0.069	**0.075**	0.016	**0.024**	0.030	**0.033**	**0.027**	0.023	0.018	**0.050**	0.010	**0.016**
Renovating home	0.179	**0.232**	0.011	**0.049**	0.071	**0.120**	0.019	**0.078**	0.085	**0.192**	0.016	**0.053**
Rock climbing	0.347	**0.375**	0.027	**0.092**	**0.217**	0.176	0.101	**0.173**	0.309	**0.322**	0.063	**0.104**
Town hall meeting	0.424	**0.456**	0.059	**0.099**	**0.270**	0.244	0.116	**0.172**	0.266	**0.379**	**0.158**	0.115
Winning race without vehicle	0.139	**0.147**	**0.082**	0.061	0.075	**0.101**	0.069	**0.081**	0.088	**0.138**	**0.073**	0.060
Working metal crafts project	0.052	**0.054**	0.019	**0.032**	0.018	**0.033**	0.020	**0.029**	0.019	**0.038**	0.020	**0.024**
Mean	0.259	**0.279**	0.067	**0.076**	0.142	**0.168**	0.082	**0.123**	0.158	**0.239**	0.063	**0.082**

when the noise amount is 30%. When 50% noise is inserted into the concept detection results for the full vocabulary, the performance drops by 11%. It means that even if 50% of the detector predictions are distorted, the event recognition performance will be degraded by only 11%. Interestingly it implies that improving the current level of concept detector accuracy has at best a limited influence on event recognition performance.

What is more, improving the detector accuracies has the same effect on event recognition performance as adding more detectors to the vocabulary. If we insert 50% noise into the vocabulary made of 50 concept detectors, for example, the event recognition performance is 0.10 in terms of MAP. We may improve the accuracy by removing the noise again, or by adding 50 more (noisy) concept detectors to the vocabulary. In both cases the event recognition performance increases to 0.13 in terms of MAP. Considering the wide availability of large amounts of training data for concept detectors [7], adding more concept detectors seems to be more straightforward than improving the detector accuracies for event recognition vocabularies.

We recommend to increase the size of the concept vocabulary rather than improving the quality of the individual detectors.

6. RECOMMENDATIONS

In this paper we study what composition of detectors in a concept vocabulary leads to the most effective event recognition in arbitrary web video. We consider four research questions related to the number, the type, the specificity and the

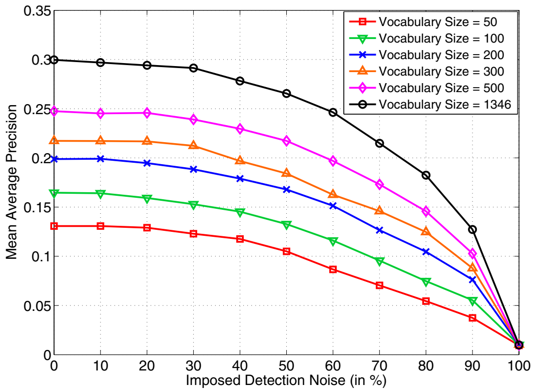

Figure 5: Experiment 4: Event recognition performance is robust when small amounts of noise are inserted into the concept detectors of the vocabulary. The more accurate the concept detectors in a vocabulary, the higher the event recognition performance. However, adding more detectors with the same noise levels may be a more straightforward way to increase event recognition performance.

quality of the detectors in concept vocabularies. From the analysis of our experiments using 1,346 concept detectors, a dataset containing 13,274 web videos, 25 event definitions, and a state-of-the-art event recognition pipeline, we arrive at the following four recommendations:

- **Recommendation 1:** Use vocabularies containing more than 200 concepts.
- **Recommendation 2:** Make the vocabulary diverse by including various concept types: *object*, *action*, *scene*, *people*, *animal* and *attributes*. However, selecting too many concepts from the same type, especially the less diverse concept types, leads to correlated concepts and should be avoided.
- **Recommendation 3:** Include both general and specific concepts into the vocabulary.
- **Recommendation 4:** Increase the size of the concept vocabulary rather than improve the quality of the individual detectors.

The recommendations may serve as guidelines to compose the appropriate concept vocabularies for future event recognition endeavors.

Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, and 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.

7. REFERENCES

[1] T. Althoff, H. Song, and T. Darrell. Detection bank: An object detection based video representation for multimedia event recognition. In *ACM MM*, 2012.

[2] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE TMM*, 4(1):68–75, 2002.

[3] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *MTAP*, 51:279–302, 2011.

[4] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li. Imagenet large scale visual recognition challenge 2011.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] A. Hauptmann, R. Yan, W. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE TMM*, 9(5):958–966, 2007.

[7] B. Huet, T. Chua, and A. Hauptmann. Large-scale multimedia data collections. *IEEE MM*, 19(3):12–14, 2012.

[8] L. Jiang, A. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, 2012.

[9] Y. Jiang. Super: towards real-time event recognition in internet videos. In *ACM ICMR*, 2012.

[10] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[11] X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *ACM ICMR*, 2011.

[12] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE TMM*, 14(1):88–101, 2012.

[13] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MM*, 13(3):86–91, 2006.

[14] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, and R. Prasad. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[15] D. Oneata, M. Douze, J. Revaud, J. Schwenninger, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. Mcguiness, S. Chen, N. O'Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, F. Basura, and T. Tuytelaars. Axes at trecvid 2012: Kis, ins, and med. In *TRECVID Workshop*, 2012.

[16] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[17] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE PAMI*, 34(5):902–917, 2012.

[18] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[19] A. Scherp, R. Jain, M. S. Kankanhalli, and V. Mezaris. Modeling, detecting, and processing events in multimedia. In *ACM MM*, 2010.

[20] J. M. Shipley and T. F. Zack, editors. *Understanding Events.* Oxford Series in Visual Cognition. Oxford University Press, 2008.

[21] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *ACM MIR*, 2006.

[22] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 43(6):76–78, 2010.

[23] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

[24] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[25] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering visual categorization with the GPU. *IEEE TMM*, 13(1):60–70, 2011.

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

[27] E. Younessian, T. Mitamura, and A. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ACM ICMR*, 2012.