Department of
Information Engineering
and Computer Science DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

(UNSEEN) EVENT RECOGNITION VIA SEMANTIC COMPOSITIONALITY

Julian Stöttinger, Jasper R. R. Uijlings, Anand K. Pandey, Nicu Sebe, and Fausto Giunchiglia

March 2012

Technical Report # DISI-12-020

Also: published at Computer Vision and Pattern Recognition Conference (CVPR), 2012.

(Unseen) Event Recognition via Semantic Compositionality

Julian Stöttinger, Jasper R. R. Uijlings, Anand K. Pandey, Nicu Sebe, and Fausto Giunchiglia
University of Trento
Trento, Italy
[julian|jrr|anand|sebe|fausto]@disi.unitn.it

Abstract

Since high-level events in images (e.g. "dinner", "motorcycle stunt", etc.) may not be directly correlated with their visual appearance, low-level visual features do not carry enough semantics to classify such events satisfactorily. This paper explores a fully compositional approach for event based image retrieval which is able to overcome this shortcoming. Furthermore, the approach is fully scalable in both adding new events and new primitives. Using the Pascal VOC 2007 dataset, our contributions are the following: (i) We apply the Faceted Analysis-Synthesis Theory (FAST) to build a hierarchy of 228 high-level events. (ii) We show that rule-based classifiers are better suited for compositional recognition of events than SVMs. In addition, rule-based classifiers provide semantically meaningful event descriptions which help bridging the semantic gap. (iii) We demonstrate that compositionality enables unseen event recognition: we can use rules learned from non-visual cues, together with object detectors to get reasonable performance on unseen event categories.

1. Introduction

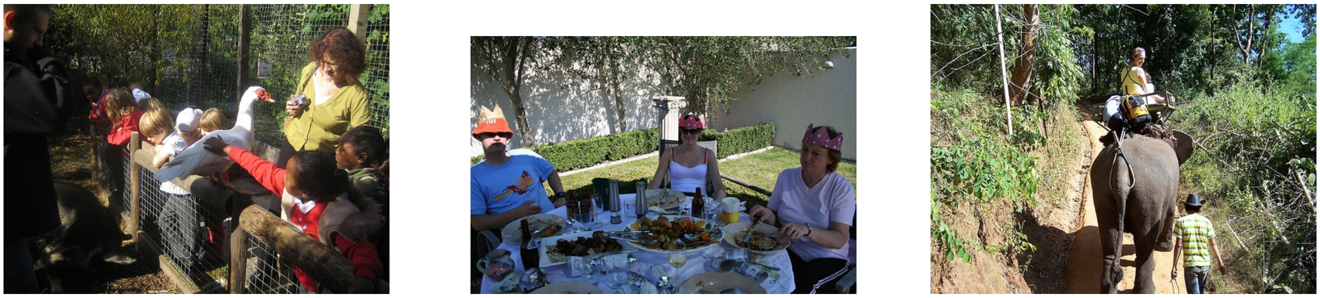
Our life is a constellation of events which, one after the other, pace our everyday activities and index our memories [4]. Events such as a birthday, a summer vacation, or a school trip are the lens through which we see and memorize our own personal experiences. In turn, global events, such as world sport championships or global natural disasters (e.g., the 2004 tsunami) or, on a smaller scale, a local festival or a soccer match, build collective experiences that allow us to share personal experiences as part of a more social phenomenon. When describing events, we ground in our experience, our common and abstract understanding of the world and the language that we use to describe it. The generic notion of "beach" is then associated to a specific time and place, which is frozen in the photo we have taken back then. The definition of an event is therefore subject to cultural and personal perspectives.

Given that knowledge is by definition incremental, in the last centuries, libraries, the keepers of knowledge, have been developing ways to cope up with the increasing amount of knowledge in a compositional way. A compositional approach to event recognition is also desirable from a computer vision perspective as the appearance may not be directly correlated with meaning: "Soccer" on grass is the same event as "soccer" on sand, yet the appearance differs.

Hence we argue that high-level semantic concepts are better recognized by their constituents. The benefits of using compositionality allow us (i) to use externally trained, general object detectors, (ii) to learn semantically meaningful rules for event classes based on object occurrences only, (iii) to learn the layout of events in images, (iv) to match semantically close images that may not be visually similar, and (v) to extend the system by attaching new events without changing the detector models. Therefore we are aiming to bridge the semantic gap by composing semantics.

For defining events, we employ an approach from Library and Information Science (LIS): the Faceted Analysis-Synthesis Theory (FAST). FAST leads to a collection of relevant events from the material to be classified. Towards this goal we have been analyzing and annotating the VOC Pascal 2007 dataset manually, leading to a faceted hierarchy of 228 high-level events. The annotation aims to describe the what, how and why as it is evident from the images: An image of the event "dinner" connotes with multiple categories as it might be a social gathering (*e.g.* barbecue), and/or a personal event (*e.g.* watching the game), and/or a festive event (*e.g.* Christmas), as shown in Figure 1. Following LIS, these events are perpetually genuine and are a subset of some kind of *universal knowledge*. It can therefore be transferred and extended to any other data-set.

Our research questions are: (i) Can we use a formal methodology to define events? (ii) Is it beneficial to use compositionality? (iii) Can we exploit compositionality to define unseen events? The paper is organized as follows. Section 2 gives the state of the art. Section 3 describes FAST to build the ground truth. In Section 4 the proposed approach is given, while Section 5 concludes.



(a) Event classes: Recreational Activity (Zoo visit)

(b) Event classes: Social Gathering (Barbecue), Get Together, Personal Event (Birthday)

(c) Event classes: Touring and Trip (Elephant Trip)

Figure 1: Example of Pascal VOC 2007 images and their FAST based high-level event annotation of 228 event classes. Note that we are aiming to recognize an elephant trip (Figure 1c) without any knowledge of an elephant.

2. State of the Art

Faceted systems are based on the idea that the universe of knowledge is incremental in nature and that characteristics are the driving force behind the semantical grouping into categories. Faceted systems are used for knowledge bases in information retrieval. They emerged in 1980 in the form of faceted thesauri to serve as a switching language to support searching across databases [1]. Faceted systems can be created using the classical FAST methodology [26] and were introduced to computer science for knowledge management in [11, 12]. Example of faceted systems are DERA [10] and the Flamenco Search Interface Project[1], which uses hierarchical faceted metadata in a manner that allows users to both refine and expand the current query while maintaining a consistent representation of the knowledge [3]. To a certain degree facet based search engines include Facetedpedia [17], query-dependent faceted interfaces for Wikipedia, FlexIR [25], a domain-specific information retrieval system that uses the notion of domain dimensions, HAYSTACK[2] and CiteSeer[3]. This paper brings the faceted system approach into the realm of computer vision.

The idea that an image can be hierarchically decomposed in its objects has been studied since the dawn of computer vision (e.g. [22]). Compositionality is often applied in object recognition. Objects are seen as a composition of their consisting sub-parts [23, 9, 33] leading to part based models of objects. Complex compositional rules can be expressed using object detection grammars [9]. By modeling the prepositions and adjectives that relate to subjective nouns, model interactions between objects are expressed in [13] . Additionally, explicitly defined spatial relationships can be incorporated in the confidence of detector results to improve their performance by taking object co-occurrence into account [7], but limit the applicability of the approach to unusual spatial set-ups. Attribute-based classification is able to transfer knowledge of object attributes across image datasets [15].

Opposed to these approaches, visual phrases [27] aim to visually learn more complex compositions of objects and actions. In certain data-sets, this decreases the size of the needed training data since the object in interactions showed to be more discriminative than the single objects. In our approach composition is driven by semantics as codified in the high level notion of events.

The most related field to this work is the field of action recognition in images. This field focuses solely on the presupposed humans in the images and analyze their poses (*e.g.* [6, 32]). A limited notion of events in image recognition is used in [18]. In their work events are defined as being a semantically meaningful human activity, taking place within selected environments. They learn the appearance of 8 activities and their background patterns which are then hierarchically combined. In all 8 activities a human and his/her pose are central to the recognition of the event.

In contrast to previous work we use a general notion of events as given in [2] based on [24]: Our life is a constellation of events which, one after the other, pace our everyday activities and index our memories [4]. Everything a picture can tell us may denote an event. Our assumption is that when we want to retrieve an image, we are most probably grounding for a representation of a certain event in our memories.

3. Building a Memorable Ground-truth

In this section we apply the Faceted Analysis-Synthesis Theory (FAST) model [26] to create a contextual mapping of events of a domain by dividing them in various *facets*. It explicitly defines events as being compositional in nature which facilitates scalability in terms of both the events and

[1] http://flamenco.berkeley.edu/pubs.html

[2] http://groups.csail.mit.edu/haystack/

[3] http://citeseerx.ist.psu.edu/

Step	Instruction
1. Defining the domain	Define: What entities are of interest to the intended user group? What aspects of those entities are of interest?
2. Formulating facets	Looking into the materials (encyclopedia, journal articles, photos, etc.) that express the interests of intended users and are useful for finding the terms related with the domain. Draw a list of candidate terms from the above sources. Sort these into homogeneous groupings (facets).
3. Structure each facet	Place the list of terms/items in a hierarchy
4. Determining the order of facets (dependent on anticipated use)	Arrange the facets into the categories. Standardize the terms with the help of a controlled vocabulary to control the semantics of the terms used.
5. Re-do step 4	If a new perspective on a domain is desired, new categories can be arranged

Table 1: Recipe how to manually build a FAST-based ground-truth hierarchy.

their primitives (*i.e.* facets). It also collocates all aspects of a domain by dividing it in fundamental categories. A change of domain, *i.e.* transfer learning, is explicitly integrated in the model. A facet can be defined as an "homogeneous group or category derived according to the process and principles of facet analysis". We may look upon these facets as groups of terms derived by taking each term and defining it, *per genus et differentiam*, with respect to its parent class [31].

The essence is the sorting of terms in a given field of knowledge into homogeneous, mutually exclusive facets, each derived from the parent universe by a single characteristic of division. Further, the facets are grouped in categories often referred as fundamental or elementary categories. Ranganathan [26] defines the fundamental categories as *personality, matter, energy, space and time*, or also as *discipline, entity, property and action*. A recipe on how to generate a hierarchy using FAST is given in Table 1. The final classification scheme can be defined as "a list of standard terms to be used in the subjective description of the documents".

We applied FAST to the *trainval* set of the Pascal VOC 2007 dataset [8], the dataset which we use in all experiments. For 3990 images we found at least one event. The rest of the images were too abstract or vague in nature to denote any event. The dataset and its full description is available online[4]. Images were analyzed to define ideas and actions such as, "child", "food", "yawning", "the zoo", and to synthesize them into an event. If there is an image containing the ideas such as, "cars" and "display" we synthesize them to articulate the event "car exhibition". This technique was applied for all images leading to 236 event classes. Many images were classified into more than one event class. To be semantically correct, we standardized the event types with standard terms to avoid any ambiguity in the intended meaning of the term (see step 4 in Table 1). The WordNet[5] database was used to find the right candidate term, *e.g.* for the event type "walking" the sense "the act of traveling by foot" was chosen. Similarly, other event types were also standardized by choosing their right terms. These event classes were not ordered and were not semantically linked with each other. FAST was applied to sort out the event types in homogeneous groups. The main characteristics group the event classes together. A characteristic is an attribute or attribute-complex which is chosen based on its semantical relevance and importance. The characteristics provide the main idea or action within the event. The root-event classes identified can be described as follows.

Personal event: An event particular to given individuals. Sub-classes include "daily routine", "eating" or "animal keeping".

Social gathering: Events celebrating or commemorating a cultural, religious, etc. occasion involving collective action rather than an individual one. Sub-classes include "dinner", "party" or "get together".

Touring and trip: Travel for pleasure. Sub-classes include "bicycle trip", "car trip", "elephant trip" and "waiting for the train".

Recreational activity: A leisure activity which refreshes and recreates. Sub-classes include "bowling" and "zoo visit".

Maintenance, repair and overhaul (MRO): The fixing of any sort of mechanical or electrical device. Sub-classes include "towing" and "building".

Natural phenomenon: A non-artificial event, *i.e.* an event not produced by humans. Sub-classes include "flooding", "snow fall" and "death".

Sport: Physical activity which aims to maintain or improve physical fitness and provides entertainment. Sub-classes include "minigolf", "rodeo" and "motorcycle racing".

Performance: Performer(s) behaving in a particular way for a group of spectators. Sub-classes include "concert", "air show" and "motorcycle slack-lining".

Exhibition: A collection of things for public display. Sub-classes include "airplane exhibition", "fair", "sheep exhibition".

[4] dataset available at http://www.feeval.org

[5] http://wordnet.princeton.edu/

4. Learning the Composition of Events

In this section we examine if the composition of events should also be reflected in machine learning. We address this question in three sets of experiments on the Pascal VOC 2007 dataset in terms of its 20 object categories: (i) How can we learn events using objects as their constituents? (ii) Can we learn events using object detectors as constituents? (iii) Can we use object detectors for unseen event recognition? All experimental results are created using 3-fold cross-validation with ten repetitions on the 118 events that have at least 3 examples.

4.1. Compositional and Visual Features

We use three types of features: (i) Ground truth object labels, yielding an upper bound on recognition through compositionality using the 20 Pascal VOC object categories only. Using the manual annotation, we are assuming perfect object detection. We use it for layout feature tuning and to show that using these constituents, which are in many cases semantically unrelated to the events, we are able to outperform the state-of-the-art in visual recognition. (ii) State-of-the-art Bag-of-Words visual features, achieving top scores in the last 3 years' Pascal challenges. This provides a baseline. (iii) Flexible object detection scores which enable fully automated recognition through the composition of object categories. Terms in italics denote their short names which they are referred to further on.

Ground Truth Features are extracted by using the ground truth (GT) object annotations of the dataset to construct the following features (compare Figure 2): (i) The *presence* or absence of each object. In this setup, the image feature vector is of a size of 20 bits only. (ii) The fraction of the image which all instances of each object category occupy (*RelSize*, feature vector length: 20). (ii) The fraction of the image which the objects occupy, using a Spatial Pyramid division. From all evaluated constellations (2,3,4 spatial pyramid, 9 and 10 quadrants in the golden ratio, etc.) 3 equal horizontal regions perform best. Every constellation of fractions dividing the image vertically decreases the performance, therefore we argue that the semantic layout of images is horizontally flip invariant. The best performing feature of three horizontal regions is referred to as *RelSize3h* providing a length of the feature vector of 60, less than half of a single SIFT vector.

Bag-of-Words Features (*BoW*) are used for comparison with state-of-the-art visual image retrieval. The implementation employs [28] using the following settings: We use pixel-wise sampling; we sample patches of 16 by 16 pixels at every possible location within the image. From these patches we extract 3 types of SIFT descriptors: SIFT, RGB-SIFT, and Opponent Color SIFT [19, 30]. We use a Random Forest as visual vocabulary [21], where we perform first PCA on the SIFT descriptors to de-correlate the dimensions and improve the discrimination of the vocabulary [28]. We use a Spatial Pyramid [16] of 1x1 and 1x3, as vertical divisions were found to reduce performance for general image classification. Finally, classification is done using a SVM with Histogram Intersection kernel using the efficient approximation of the classification function of [20].

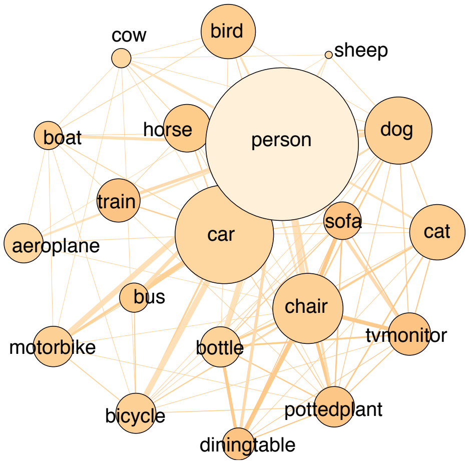


Figure 2: Graph of occurrences of objects in Pascal VOC 2007, strength of edges indicate the number of co-occurence. size of nodes indicates the number of occurrences.

Great advances in *exact* object localization have been made recently (*e.g.* [9]). Unfortunately, many images do not show the whole objects. Additionally, typical environments of objects make the object itself more likely to be present in the image. For example, a horse barn and meadow make an image *horsey*, even if there is no horse present. We pick up an exciting new idea from [29] to perform a relaxed object localization using the **Most Telling Window Features** (*MTW*). It provides that window within the image which is most discriminative for the object class. This can be a complete object, but also an object part, or a collection of objects. By focusing on the most discriminative part instead of the exact object boundaries, scores for classification, determining if the object is present in the image or not, are higher than for exact object localization. Moreover, by using localised features we avoid confusion which may be caused by the background as well as double counting the background when more object categories are taken into account.

The Most Telling Window uses the same Bag-of-Words features, vocabulary, and classifier as described earlier. During testing, instead of using the whole image, we evaluate multiple windows and select the one with the highest classification score. We use a Spatial Pyramid of 2x2 for each window. For training, for each object category we use its ground truth bounding boxes as positive learning examples. Ground truth bounding boxes for objects of a different class serve as negative learning examples. We mine difficult negative images by selecting the highest scoring window of each negative image of the training set and retrain the classifier. We repeat this two times after which performance stabilizes. The output of the MTW yields a single score for each Pascal VOC object category, resulting in a feature vector of length 20.

4.2. Finding Conditions for Event-identity

We want to find conditions for determining the identity of an event in terms of its constituents, which are the 20 Pascal object categories in this paper. The event-identity is based on the *non-duplication principle* [2]. There are two main theories: Davidson [5] defines it based on the sameness of causes and effects, Quine [24] aims for the unique spatio-temporal location. The common objective is to discover sufficient conditions for event-identity, *i.e*. find a function R for which it is true and not trivial that for any event e_1 and for any e_2,

$$R(e_1, e_2) \rightarrow e_1 = e_2. \qquad (1)$$

The problem is to discover values of R that make this true but not trivial. No two events can be related by $R(x, y)$ having the very same causes and effects, and are not the same events. Therefore, in the classification of events, we face the problem of individuation of events, i.e. finding a solvable and satisfying R, keeping in mind that we do not know the complete where, how and when a given image is taken. In the following, we describe how we aim to find the rule-set R, the sufficient conditions for event-identity. We perform this using first Support Vector Machines and then building a decision tree.

We use **Support Vector Machines** (SVM) as a baseline for learning R. Note that this leads to semantically not interpretative rules for R. We use the well-known libsvm[6] library. Experiments are carried out using RBF, Histogram Intersection, and χ^2-kernel. In the compositionality experiments, the χ^2-kernel outperformed other kernels. Therefore we only report the scores of the χ^2-kernel in the experimental section.

We aim to find sufficient conditions for R using **rule-based classification** to determine the identity of an event in the image. Our goal is a formal system based on the occurrences of the primitives of the training set. Let every of the m images belonging to one event be described by the conjunction of the existence of the primitives $p_{1..n}$. Having a sufficiently large training set of j instances, we then assume that any event e is modeled by

$$e \models \bigvee_{j=1}^{m} \bigwedge_{i=1}^{n} p_{i,j}. \qquad (2)$$

This leads to one expression in full disjunctive normal form (DNF) per event describing the training set. The canonical DNF is derived in linear time using a linear hash map to remove equal and dual terms. Since this is done independently per event, a new event can be added to the formal system by simply adding the new expression to the prior set. Note that annotated objects are not necessary. Training can also take place based on unsupervised detector results.

By learning a pruned decision tree of all e, we derive a minimal R, which is possibly overlapping with other events, with the best trade-off between description length and classification accuracy by cross-fold validation. Some events are indistinguishable using the primitives in the Pascal VOC set. For these events their probability is defined by their relative occurrence in the training set.

4.3. Learning Events through Objects

In this section we investigate how we can best learn events using the objects contained within the image. We feed the proposed features into a SVM with χ^2 kernel. Results are presented in Table 2. As it can be seen, using the presence of objects based on the ground truth gives a MAP of 0.178, which is significantly higher than the random result of 0.022 MAP. Using the relative size of the objects in the image we achieve a MAP of 0.245. Using the best performing layout with three horizontal divisions, MAP rises to 0.258 MAP.

However, as events are compositional in nature we hypothesize that an SVM may not be the best learning method at this stage and instead a knowledge or language based system could perform better. Figure 3 gives an example. It provides the rules for three events in minimal product-of-sums form (PoS) [14] to make the expressions e easier to grasp. The events are of similar constituents: Motorbike exhibition, motorbike racing and motorbike stunts. All 3 event classes provide predominantly persons on motorbikes. For a human observer it is straightforward to determine the differences: Motorcycle exhibitions take normally place in a convention center, where people are admiring highly polished motorbikes. Motorcycle racing events are predominantly determined by common vest numbers of the drivers and race track lanes. Motorcycle stunt images are typically showing pictures of people doing crazy things on their motorcycle. Since we only have a person and a motorcycle detector we are not able to find a perfect R. However, some insights can be derived. The data-set provides some beach rally pictures, which means that sometimes, there is a boat seen in the event. This does not apply to the other events, as visualized in Figure 3. In contrast, exhibition pictures do

[6] http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm/

(a) exhibition $\models \text{common} \wedge \neg \text{boat} \wedge (\neg \text{car} \vee \text{person})$

(b) racing $\models \text{common} \wedge (\neg \text{boat} \vee \neg \text{bus}) \wedge (\neg \text{boat} \vee \neg \text{car}) \wedge (\neg \text{boat} \vee \text{person}) \wedge (\neg \text{bus} \vee \neg \text{car}) \wedge (\neg \text{bus} \vee \text{person}) \wedge (\neg \text{car} \vee \text{person})$

(c) stunt $\models \text{common} \wedge \neg \text{boat} \wedge \text{person}$

Figure 3: Visual examples and minimal PoS form of motorcycle events: (a), (b) and (c) share a common term of occurrence of motorbikes and the absence of most other primitives. They differ only in their composition of boat, bus, car and person

	baseline SVM	GT & **SVM**	GT & **tree**	GT, size & tree	GT, layout & tree	SIFT & SVM	MTW & SVM	MTW & tree
Selected events	**Random**	**Presence**	**Presence**	**RelSize**	**RelSize3h**	**BoW**	**MTW**	**Unseen**
Air show	0.015	0.293	0.286	0.488	0.447	**0.654**	0.497	0.281
Car trip	0.015	0.069	**0.097**	0.091	0.142	0.076	0.086	0.065
Cat Play	0.016	0.347	**0.452**	0.148	0.178	0.097	0.083	0.448
Dinner	0.020	0.342	0.478	0.498	**0.536**	0.274	0.312	0.093
Exhibition	0.048	0.176	0.221	0.235	0.278	**0.391**	0.256	0.136
Festival	0.023	0.072	0.102	0.066	0.079	**0.284**	0.086	0.024
Motorcycle Stunt	0.004	0.160	0.107	0.148	**0.442**	0.034	0.042	0.226
Rodeo	0.014	0.667	0.825	**0.838**	0.839	0.311	0.287	0.008
Walking the dog	0.011	0.184	0.327	**0.418**	0.342	0.118	0.080	0.044
MAP all events	**0.022**	**0.178**	**0.242**	**0.245**	**0.258**	**0.199**	**0.177**	**0.120**

Table 2: Overview results of event based image retrieval in the VOC Pascal 2007 (118 eventclasses found in 5011 images. 10124 event annotations in total)

not necessary need a person in the picture. And sometimes, even a car appears. There is no legit stunt without riding the bike, therefore there is always a person in the motorcycle stunt events. This confirms that there is more information in the presence information of the 20 object categories than the results of the SVM imply.

We therefore turn to classification using logical decision rules. An effective and common prediction tool dealing with such rules is a decision tree. Thus we use a decision tree as classifier on the presence of the 20 object categories (still using the ground truth). To make each split in the decision tree, we consider each object separately and choose the one whose presence results in the largest information gain defined in terms of Shannon Entropy over the event categories. The results are shown in Figure 2. As can be seen, for the decision tree the MAP is 0.242, far higher than the results of the SVM (0.178) and on par with layout features. This suggests that a logical or tree based based representation may be better for a classification of events based on their constituents, the objects present in the image.

Another major advantage of using a decision tree based on objects is that we can now give logical rules defining how to define an event. For example, Figure 4 shows a decision tree learned of one of the folds. The decision tree yields a natural ordering of which objects are significant for the events. In terms of the FAST analysis it gives an idea of the *characteristic* of the events. We see that the presence or absence of a *person* is the most important, which is intuitive as our events arise from our human-centered view of the world.

Cars are the second important as it is the second most occurring object in this dataset, which may be a bias of the dataset but also reflects its importance in the western world (there are just many photos of cars on the internet in general). Note that by its dominance a car is a powerful indoor/outdoor indicator.

To appreciate how good such MAP scores are in a retrieval task, consider the following. Giving one example in a fold (there are few examples per event category), a score of 0.20 MAP means that the target image occupies on av-

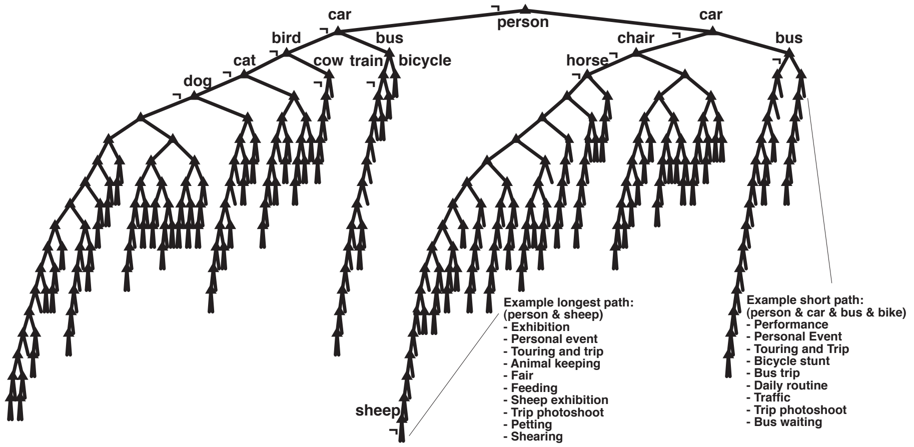


Figure 4: Decision tree which visualizes the learned boolean expressions for event recognition. Each left branch denotes absence, each right branch presence of a primitive. Most significant primitives are labeled on top of the tree, the least significant one is labeled at the bottom. For one short and for the longest path, the possible events are shown in order of their occurrence probability within the node.

erage the fifth position of the ranking. This is a promising result for difficult retrieval tasks, especially since the false positives are not visually similar, but are semantically close (compare ambiguities in Figure 4).

4.4. Learning Events through Object Detectors

In this experiment we use the Most Telling Window classification method to yield a score for each of the 20 objects. We use these scores as input for a Support Vector Machine in the same manner as in the previous section. Using the 20 object detection scores, we are able to get a MAP of 0.177. This is quite impressive considering that the upper bound using only these 20 categories is 0.178 (without using layout information). The use of the 20 object detectors already outperforms global Bag-of-Words for some event categories such as Personal event, Aeroplane exhibition, Bicycle trip, Boating, Bus trip, Couching, Dinner, Motorcycle exhibition, Road trip, Train exhibition, Train trip, Trip photoshoot, Voyage, Wedding, etc. All these events convey a higher level semantics.

To summarize, our results show that it makes sense to recognize events by their constituents. For many event categories of a higher semantic level our use of only 20 object categories already enables us to outperform state-of-the-art Bag-of-Words. Note that this 20 dimensional object representation feature vector is negligibly small compared to today's Bag-of-Words representations.

4.5. Unseen Event Recognition

In the final experiment we test if we can do *unseen* event recognition. The basic idea is that one can learn from any domain what are the constituents of an event. We now use the presence/absence within the images. However, text or handcrafted knowledge can also be used. In this experiment, we learn a binary decision tree using the presence of the 20 objects within one fold of the dataset. Hence, this decision tree is learned without using any image features: We use only the MTW detector results to determine which objects are present within the image. This is done by binarizing the output: a negative distance to the decision boundary means that an object is absent. A positive distance means that the object is present. In this sense, we can categorize images into event classes using visual information without having learned these event classes from visual information. The results are shown in Table 2.

The performance of 0.120 MAP is about half of using the ground truth presence. In terms of search results, it means that if there is one true positive example, we retrieve it on average in the top ten results. This proves first of all that our object detectors are good enough for constituent based event recognition. Secondly, for learning we use only the high-level terms of events and externally trained detectors. In this sense, no visual features are used to directly learn the appearance of events. Instead we get a compositional event recognition approach, which enables reuse of individual visual object classifiers and which facilitates the addition of extra event categories without (re-)training visual classifiers.

5. Conclusion

Our solution towards bridging the semantic gap is to use humans to provide semantics. In terms of an event hierarchy, built according to the FAST methodology, events provide the semantic glue which allows to compose the results of prior visual analysis. The main idea is that these events are not task or domain specific, but are settled once and for all, similar to what library scientists do. A library system never becomes wrong when new knowledge is being added.

This paper shows that logical representations often used in knowledge based representations have a large potential: (i) They allow to recognize events based on the presence of the objects, a logical decision tree yielded 0.242 MAP, where an SVM yields a score of 0.178 MAP. (ii) A decision tree yields human interpretative rules, which helps in breaching the semantic gap. Since we generate semantically meaningful rules, there is no reason why these rules cannot be generated from text or even be handcrafted by the user. In the latter case, this means that our system will be able to perform reasonably well when an event is either queried and/or learned as textual description. (iii) Representing events through their constituents yields a highly flexible framework for event-based image retrieval. In particular we showed that we can do unseen event recognition by using only 20 object detectors with a reasonable retrieval rate for 118 event classes of 0.120 MAP.

Acknowledgments

This work was partially supported by the European Commission under contract FP7-248984 GLOCAL and FP7-287704 CUBRIK.

References

[1] J. Aitchison. Integration of thesauri in the social sciences. *International Classification*, 8(2):75–85, 1981. 2

[2] J. Bennett. *What Events Are*. Clarendon, 1988. 2, 5

[3] V. Broughton and A. Slavic. Building a faceted classification for the humanities: principles and procedures. *Journal of Documentation*, 63(5):727–754, 2007. 2

[4] N. R. Brown. On the prevalence of event clusters in autobiographical memory. *Social Cognition*, 23(1):35–69, 2005. 1, 2

[5] D. Davidson. The individuation of events. *Essays on Actions and Events*, 1(9):163–181. 5

[6] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 2

[7] C. Desai, D. Ramanan, C. Fowlkes, and U. C. Irvine. Discriminative models for multi-class object layout. In *ICCV*, 2009. 2

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 3

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32:1627–1645, 2010. 2, 4

[10] F. Giunchiglia and B. Dutta. DERA: A faceted knowledge organization framework. Technical report, KnowDive, DISI, University of Trento, 2010. 2

[11] F. Giunchiglia, B. Dutta, and V. Maltese. Faceted lightweight ontologies. In *Conceptual Modeling: Foundations and Applications*, 2009. 2

[12] F. Giunchiglia, V. Maltese, and B. Dutta. Domains and context: first steps towards managing diversity in knowledge. *Journal of Web Semantics*, 12, 2012. 2

[13] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, pages 16–29, 2008. 2

[14] F. J. Hill and G. R. Peterson. *Introduction to Switching Theory and Logical Design*. Wiley & Sons, 1974. 5

[15] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 4

[17] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia. In *WWW*, pages 651–660, 2010. 2

[18] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. *ICCV*, pages 1–8, 2007. 2

[19] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60:91–110, 2004. 4

[20] S. Maji, A. C. Berg, and J. Malik. Classification using Intersection Kernel Support Vector Machines is Efficient. In *CVPR*, 2008. 4

[21] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, pages 985–992, 2006. 4

[22] Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *ICPR*, pages 752–754, 1978. 2

[23] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. *CVPR*, pages 1–8, 2007. 2

[24] W. Quine. *Word and Object*. The MIT Press, 1960. 2, 5

[25] S. Radhouani, C.-L. M. Jiang, and G. Falquet. Flexir: a domain-specific information retrieval system. *POLIBITS*, 39:27–31, 2009. 2

[26] S. R. Ranganathan. *Prolegomena to Library Classification*. Asia Publishing House, 1967. 2, 3

[27] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. *CVPR*, pages 1745–1752, 2011. 2

[28] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-time Visual Concept Classification. *TMM*, 2010. 4

[29] J. R. R. Uijlings, K. van de Sande, A. Smeulders, T. Gevers, N. Sebe, and C. Snoek. Most telling window. In *The PASCAL Visual Object Classes Challenge Workshop*, 2011. 4

[30] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *TPAMI*, 2010. 4

[31] B. Vickery. Faceted classification schemes. *Systems for the Intellectual Organization of Information*, 5, 1966. 3

[32] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 2

[33] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010. 2