TRECVid Semantic Indexing of Video: A 6-Year Retrospective

George Awad, Cees G M Snoek, Alan F Smeaton, Georges Quénot

Paper

TRECVid Semantic Indexing of Video: A 6-Year Retrospective

George Awad [†1], Cees G. M. Snoek [†2], Alan F. Smeaton [†3], Georges Quénot [†4]

Abstract Semantic indexing, or assigning semantic tags to video samples, is a key component for content-based access to video documents and collections. The Semantic Indexing task has been run at TRECVid from 2010 to 2015 with the support of NIST and the Quaero project. As with the previous High-Level Feature detection task which ran from 2002 to 2009, the semantic indexing task aims at evaluating methods and systems for detecting visual, auditory or multi-modal concepts in video shots. In addition to the main semantic indexing task, four secondary tasks were proposed namely the "localization" task, the "concept pair" task, the "no annotation" task, and the "progress" task. It attracted over 40 research teams during its running period.

The task was conducted using a total of 1 400 hours of video data drawn from Internet Archive videos with Creative Commons licenses gathered by NIST. 200 hours of new test data was made available each year plus 200 more as development data in 2010. The number of target concepts to be detected started from 130 in 2010 and was extended to 346 in 2011. Both the increase in the volume of video data and in the number of target concepts favored the development of generic and scalable methods. Over 8 million shots×concepts direct annotations plus over 20 million indirect ones were produced by the participants and the Quaero project on a total of 800 hours of development data.

Significant progress was accomplished during the period as this was accurately measured in the context of the progress task but also from some of the participants' contrast experiments. This paper describes the data, protocol and metrics used for the main and the secondary tasks, the results obtained and the main approaches used by participants.

Key words: TRECVid, video, semantic indexing, concept detection, benchmark.

1. Introduction

The TREC conference series has been sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies since 1991. The goal of the conference series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent annual evaluation (TRECVID) with a workshop taking place just before TREC[1)*]. During the last 15 years of operation, TRECVid has addressed benchmarking of many component technologies used in video analysis, summarisation and retrieval, all with the common theme that they are based on video content. These include shot boundary detection, semantic indexing, interactive retrieval, instance retrieval, and ad hoc retrieval, rushes summarisation, and others.

From 2002 to 2009 inclusive, TRECVid included a task on detection of "High Level Features" (HLFs), also known as "semantic concepts"[2)]. In 2010, this task evolved as the "Semantic Indexing" (SIN) task. Its goal is similar; assigning semantic tags to video shots, but it is more focused toward generic methods and large scale and structured concept sets. A more general and varied type of data has been collected by NIST than had been used in previous years of TRECVid which was split into several slices constituting the training and/or

Received ; Revised ; Accepted

†1 Dakota Consulting, Inc; NIST (Silver Spring, MD, USA)
†2 University of Amsterdam (Amsterdam, The Netherlands)
†3 Dublin City University (Dublin, Ireland)
†4 Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France CNRS, LIG, F-38000 Grenoble, France (Grenoble, France)

* http://trecvid.nist.gov/

testing sets for the 2010 to 2015 issues of the SIN task.

The SIN task has gradually evolved over the period of its running, both in the number of target concepts and the data set sizes. Also, besides the main (or primary) concept detection task, several variants of the task (or secondary tasks) have been run, including a "concept pair" task, a "localization" task, a "no annotation" task, and a "progress" task. As with the earlier HLF detection task, the indexed units in the SIN task are video shots, not full video documents.

The semantic indexing task is related to the Pascal Visual Object Classification (VOC)[3], ILSVRC[4] and other benchmarking tasks whose goal is to automatically assign semantic tags to still images. The purpose of this paper is to gather together the major contributions and to identify trends across the 6 years of the semantic indexing track and its variations. The paper is organized as follows: section 2 describes the data used for the semantic indexing task, its origin and organisation; section 3 describes the metrics used in TRECVid for evaluation; section 4 describes the main concept detection task and the results achieved across participating groups; sections 5, 6, 7, and 8 describe the concept pair, localization, no annotation and progress secondary tasks respectively. Each task description includes a short overview of the methods used by various participants.

This overview paper does not intend to be exhaustive or an in-depth summary of all the approaches taken by all the participants in all the 6 years of the running of the SIN task. Instead, it aims at illustrating the progress achieved over the period through a number of selected contributions. Full details of all the work done in the task, approaches taken and results achieved, can be found in the annual workshop proceedings, available on the TRECVid website*.

2. Data

2.1 IACC collections

In 2010, NIST collected a new set of internet videos (referred to in what follows as IACC, standing for Internet Archive Creative Commons) characterized by a high degree of diversity in creator, content, style, production qualities, original collection device/encoding, language, etc., as is commonly found in much "Web video". The collection also has associated keywords and descriptions provided by the video donor. The videos are available under Creative Commons (CC) licenses** from the Internet Archive (IA)***. The only selection criteria imposed by TRECVid beyond the Creative Commons licensing is one of video duration where the videos were required to be less than 6.4 min in duration. Seven slices (or sub-collections) of about 200 hours of video each have been created. These are officially labeled: IACC.1.tv10.training, IACC.1.A-C, and IACC.2.A-C and described in Table 1.

As can be seen, not all the slices or video sub-collections have been selected in the same way:

IACC.1.A-C have been selected as the shortest videos up to a duration of about 3.5 minutes (211 seconds) and split into three slices (A, B and C) in a symmetric way by interlacing the list, sorted by video length.

IACC.1.tv10.training has been selected as the subsequent 200 hours among the next shortest videos, up to about 4.1 minutes in duration.

IACC.2.A-C have been selected as the subsequent 600 hours of the next shortest videos up to about 6.4 minutes in duration, and then split into three slices (A, B and C) in a symmetric way by interlacing the list, sorted by video length. These include a few videos shorter than 4.1 minutes as these had been included into the global IACC collection subsequently.

Table 1 also indicates which video collection slices were used for system training and which were used for system evaluation (testing) for each year of the SIN task. From years 2011 to 2013 included, a new slice was introduced each year as "fresh data" for year N while both the test and training data from year $N-1$ was merged to become training data for year $N-1$. From years 2013 to years 2015 included, the training data (as well as the annotations) were frozen so that the "progress" task (described in section 8) could be conducted properly. While the IACC.2.A-C slices were used as test collections for years 2013-2015 respectively as "fresh data", they were made available after 2013 so that participants could provide anticipated and blind submissions for years 2014 and 2015 with their 2013 systems and anticipated and blind submissions for year 2015 with their 2014 systems.

2.2 Master (reference) shot segmentation

As in the earlier HLF task, a common shot segmentation was provided to participants so that they could

* http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

** https://creativecommons.org/licenses/

*** https://archive.org/

Table 1 IACC collections statistics

Collection (slice)	*total duration (h)*	*video files*	*min/mean/max duration (s)*	*video shots*	*mean duration (s)*	used for training	used for test
IACC.1.tv10.training	198	3,127	211/228/248	118,205	6.04	2010-2015	-
IACC.1.A	220	8,358	11/95/211	144,757	5.48	2011-2015	2010
IACC.1.B	218	8,216	11/96/211	137,327	5.72	2012-2015	2011
IACC.1.C	221	8,263	11/96/211	145,634	5.46	2013-2015	2012
IACC.2.A	199	2,407	10/297/387	110,947	6.46	-	2013
IACC.2.B	197	2,368	10/299/387	106,611	6.65	-	2013-2014
IACC.2.C	199	2,395	10/298/387	113,046	6.32	-	2013-2015
Total	1452	35,134	10/149/387	876,527	5.97	N.A.	N.A.

make submissions in the same way and so that evaluation could be made consistently across submissions using a standard information retrieval procedure. The shot segmentation was performed using an improved version of the Laboratoire dinformatique de Grenoble (LIG)[*4] tool evaluated in the TRECVid 2006 shot boundary detection task. This tool has a good detection rate, especially for gradual transitions[5]. Errors in shot boundary detection are not as critical for the concept detection evaluation as for the main search task, and in the concept pair variant participants were only asked to tell whether a target concept is visible, or not, at least at some point within a given video. A separate task has been defined for the evaluation of the temporal and spatial localization of target concepts described in section 6.

The reference segmentation of video into shots is given in several formats, including simple frame numbers in a text file and an MPEG-7 version[*], the latter being the official reference. One MPEG-7 file is provided for each video file of each (sub-)collection (or slice). Additionally, for each issue (from 2010 to 2015), an XML file specifies the list of files that should be used for training and for testing.

2.3 Key frames

A reference key frame has also been selected for each video shot and the locations of these key frames are included in the segmentation files. In order to select the best key frame within each shot, three criteria were used: (i) closeness to the center of the shot, in order to avoid gradual transition regions if any, (ii) slow motion in the neighborhood of the frame, in order to avoid fuzzy contents, and (iii) high contrast, for having a clean content representation. All these criteria were computed on each video frame using a simple and ad hoc metric. The corresponding scores were then normalized and and averaged. The frame within a shot having the highest score was selected. Archives with the extracted key frames were also made available to participants though SIN detection methods which use the whole shot rather than just the key frames has become the norm in TRECVid and elsewhere were encouraged.

2.4 Speech transcription

Speech transcription of the audio track was generously contributed by the Laboratoire d'Informatique pour la Mcanique et les Sciences de l'Ingnieur (LIMSI)[**] laboratory using their large vocabulary continuous speech recognition system[6]. In practice, the IACC collection is highly multi-lingual (tens of different spoken languages are mentioned in the meta-data) and many files also include speech in different languages. Many files did not include audio or included audio but no speech. The LIMSI transcription process was therefore conducted in two steps. In the first (for the files in which audio and speech were present) they applied an automatic language detection system. Then, when they detected a language for which they had an automatic speech transcription system, they produced a transcription, otherwise they applied by default their English transcription system. This latter choice is sensible because even if the actual language spoken in the video is different, it may still include English words, especially for technical terms, and proper nouns may also be recognized if pronounced in a similar way.

2.5 Target concept set

A list of 500 target concepts was generated, 346 of which have been collaboratively annotated by TRECVid participants (see section 2.6). The target concepts were selected as follows. First, they were chosen so that they include all the TRECVid HLFs from 2005 to 2009 in order to permit cross-collection experiments. Second, they also include the CU-VIREO374 concept set[7] which was also widely used in previous TRECVid experiments as a subset of the annotated part of the Large Scale Concept Ontology for Multimedia (LSCOM)[8]. All of these concepts were already

[*4] https://www.liglab.fr/

[*] http://mpeg.telecomitalialab.com/working_documents.htm#MPEG-7

[**] https://www.limsi.fr/en/research/tlp

selected using a number of criteria among which: expected usefulness in a content-based video search system, coverage and diversity. This set was then completed by additional concepts selected among the 3000 available in the last version of LSCOM to which a few were specifically added. The added concepts were selected in order to improve the coverage and diversity of the set as well as for creating a number of generic-specific relations among the concepts. Considering diversity, we specifically managed to have a significant number of samples for the following concept types (not exhaustive): humans, animals, vehicles, scenes, objects, actions and multi-modal (involving audio). The structure of the concept set was enriched with two relations, namely *implies* and *excludes*. The goal here was to promote research on methods for indexing many concepts and subsequently using ontology relations between them to enhance the accuracy of concept detection.

The list of the 500 TRECVid SIN concepts is available on the TRECVid web site *. Each concept comes with a TRECVid SIN identifier, the corresponding LSCOM identifier, a name and a definition. In addition, the correspondence with previous TRECVid HLF identifiers and with concept definitions in other benchmarks (e.g. Pascal VOC) are also given when available in order to facilitate cross-collection experiments.

2.6 Collaborative annotation

As most concept detection methods rely on a supervised learning approach, it was necessary to create annotations for the training of participants' systems. As no funding was initially available for this annotation process and as for the 2003-2009 HLF tasks, participants themselves were involved in the annotation process, each of them contributing at least 3% of the target volume while receiving, in return, the full set of annotations. Some funding from the Quaero project** later helped to increase the volume of annotations.

The set of target concepts and the set of training video shots were both large and as a consequence, only a fraction of the training set could be annotated, even using the "crowd" of TRECVid SIN participants and with Quaero support. Also, as most of the target concepts were sparse or very sparse in the training collection (less or much less than 1%), an active learning procedure was used in order to prioritize annotations of the most useful sample shots[9].

A system with a web interface was provided to participants for producing their annotations. They were required to annotate one concept at a time for a set of video shots represented by their reference key frames. If the key frame alone was not sufficient to enable making a good decision, they could play the full video shot. For each (concept, shot) combination, they had to choose a label as either *positive* (the concept is visible in the shot), *negative* (the concept is not visible in the shot), or *skipped* (ambiguous or bad example).

In addition to the active learning to select shots for annotation, an active cleaning procedure was included in the annotation system. Its aim was to improve the annotation quality by asking for a "second opinion" when manual annotations strongly disagreed with a prediction made by cross-validation from other available annotations. A second opinion was also systematically asked for all *positive* and *skipped* annotations as these were quite rare and their correction were likely to have a significant impact[10]. In case of disagreement between the first and second opinions, a third opinion was asked for and a majority vote was applied. The system enforced that second and third opinions were asked of different annotators. The annotation system also made use of the provided set of relations in order to increase the number of annotations and to enforce a consistency among them. In the last version of the collaborative annotation, 8,158,517 annotations were made directly by the participants or by the Quaero annotators and a total of 28,864,844 was obtained by propagating those initial annotations using the *implies* or *excludes* relations.

In order to improve annotation efficiency and as was done in the years from 2011 to 2013, the test set of year $N-1$ was included in the development set of year N, the assessments on year $N-1$ as well as the participants' systems' outputs on year $N-1$ were all used to bootstrap the active learning for the additional annotations produced for year N. For each year from 2010 to 2013 a new set of annotations was performed and added to the global pool.

3. Metrics

For the semantic indexing or concept detection task, the progress task and the concept pair task, the official TRECVid metric is the Mean Average Precision (MAP) which is a classic metric in information retrieval.

* http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls/

** http://www.quaero.org/

Table 2 2010-2015 SIN tasks summary

	Data		Number of Concepts			Secondary Tasks			
Year	Training data	Test data	Annotated concepts	Submitted concepts	Evaluated concepts	Concept pairs	Localization	No annotation	Progress
2010	IACC.1.tv10.training	IACC.1.A	130	10/130	10/30	-	-	-	-
2011	2010 train + 2010 test	IACC.1.B	346	50/346	23/50	-	-	-	-
2012	2011 train + 2011 test	IACC.1.C	346	50/346	15/46	10	-	Yes	-
2013	2012 train + 2012 test	IACC.2.A	346	60	38	10	10	Yes	Yes
2014	2013 train	IACC.2.B	346	60	30	-	10	Yes	Yes
2015	2013 train	IACC.2.C	346	60	30	-	10	Yes	Yes

In practice, however, MAP is evaluated on a statistical basis using the Inferred[11] and Extended Inferred[12] Mean Average Precision method using the `sample_eval` tool[***] available from the TRECVid web site. Evaluation is based on an assessment of a subset of the test set built by pooling the top of the submissions from all participants. Additionally, in the Inferred Average Precision (InfAP) approach, the pools are split into sub-pools, some of which are only partially assessed, the first sub-pool being 100% assessed and the following sub-pools being more and more sub-sampled. The extended InfAP approach correspond to a further improvement in the estimation method. The main goal of the inferred approach is to estimate the MAP value with a good accuracy while using much less assessments. In practice, we used it in order evaluate more concepts (typically twice as many) for the amount of manpower that was allocated for assessments. While doing this, we remained conservative in the pool partitioning and in the selection of the corresponding sub-sampling rates. We also conducted experiments using the submissions of previous years for which the whole pools were assessed at 100% and checked that (i) the inferred MAP values were very close to the actual ones, (ii) the ranking of the systems was not changed.

4. Concept detection task

4.1 Task definition

The task of automatic concept detection from video is defined as follows:

> "Given the test collection, master shot reference, and concept definitions, return for each target concept a list of at most 2000 shot IDs from the test collection, ranked according to their likelihood of containing the target."

The training conditions, data and annotations are not part of the task definition. However, participant submission types are defined according to the following:

- **A** used only IACC training data;
- **B** used only non-IACC training data;
- **C** used both IACC and non-IACC TRECVid (Sound & vision (S&V) and/or Broadcast news) training data;
- **D** used both IACC and non-IACC non-TRECVid training data;
- **E** used only training data collected automatically using only the concepts' name and definition;
- **F** used only training data collected automatically using a query built manually from the concepts' name and definition.

Type A corresponds to using only the official training data (as presented in Table 1) and the corresponding collaborative annotation (described earlier in section 2.6). Type D corresponds to using whatever training data is available. Types E and F have been added in order to encourage research on systems able to work without prior annotation based on including an automatic crawling tool instead (described later in section 7).

Table 2 gives an overview of the specifics of the 2010 to 2015 issues of the SIN task. The first part of the table indicates what training (for type A submissions) and testing data were used in each year. The second part indicate the number of concepts for which annotations were provided, the number of concepts for which participants were required to submit results, and the number of concepts that were actually evaluated. From 2010 to 2012 included, two versions of the task, "light" and "full", were proposed to participants. The numbers are displayed as light/full in these cases. The third part of the table indicates which of the secondary tasks were available for the different years and the corresponding number of targets for those secondary tasks, if relevant. These secondary tasks are described in sections 5, 6, 7 and 8.

From 2010 to 2012, we attempted to scale up the task in order to encourage the development of scalable methods and to follow the ImageNet and LSCOM trends to

[***] `http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/`

increase the number of target concepts. Meanwhile, we also offered a light version of the task so that teams unable to follow the increase in the number of concepts could still participate and so that advanced but not yet scalable methods could also be evaluated. Considering participants' feedback, we froze the concepts set size to 346 concepts from 2011 onward. Also, since 2013, considering that the 2010 to 2012 results were consistent between the light and full submissions for the participants that made both, we removed the light/full distinction, replacing it by a single intermediate "main" task with 60 target concepts. These are a subset of the previous 346 "full" concept set and, even though submissions were required only for the 60 concepts in the "main" set, annotations were still made available for the full set and many participants actually computed for the full set and submitted only on the main set.

4. 2 Results

Figures 1 to 6 show the performance obtained by the SIN task participants for the 2010 to 2015 issues of the task respectively. Participants were allowed to make up to four A- to D-type submissions (not necessarily one from each type) plus two additional E- or F-type submissions when possible. For simplifying the visualization, we display on the plots only the best submission from each participant for each submission type. Participants were required to define a priority among their submissions according to their prediction of which one would be the best-performing but we selected here only the actual best one for each submission type. As some participants made submissions with different types, those participants appear several times in the plots.

The total number of participants were respectively 39, 28, 25, 26, 15 and 15 for the 2010 to 2015 issues of the SIN task. From 2010 to 2012, the numbers included participants to both the light and full versions of the SIN task. However, as the light concept set was included in the full one, all submissions for the full task were added to the submissions for the light task. Respectively, 28, 18 and 15 participants made submissions to the full task only in the 2010, 2011 and 2012 issues.

As the test collections and the concepts selected for evaluation differed each year, it is not possible to compare directly the MAP performances across the different issues of the task. The increase of best and median MAP values from 2010 to 2013 is probably partly related to improvements in the methods but it is also likely related to differences in the intrinsic difficulty of the task because of the nature of the video used, and the concepts selected. The size of the training set and the number of available annotations also significantly increased during the years of the task which was the motivation for the introduction of the progress secondary task over the 2013-2015 period (section 8).

Similarly, for the 2010-2012 issues, even though the test collection used is the same, it is not possible to compare directly the MAP performances between the light and full tasks as the concept sets are different. However, it is possible to compare the ranking among systems (or participating teams) that submitted results to the full task (which also appear in the light task), by filtering the submission to the smallest concept set. We can observe that these system/participant rankings are quite consistent across the two versions of the task and even though there are some permutations, there are quite few of them and when they happen the performances of the involved systems are quite comparable. This good stability observed on the 2010-2012 issues validated the choice of keeping only a concept set of intermediate size (60).

For simplicity and for ease of comparison, we display all the submissions for the same year/task in a single graph. However, it should be noted that fair comparisons between approaches should in principle be made only among submissions of the same type (even within a same year/task). Differences in submission types correspond to different training conditions, the main difference being that some actually use more training data or different training data than others, possibly with similar methods. The difference is especially important between the A-D types that use explicitly and purposely annotated data and the E-F types that do not but use instead only data gathered via general search engines which return noisy results that are not manually checked or corrected.

Figure 7 shows the per concept InfAP for the 2015 main task. Results are very similar for the other years. It can be observed that while the MAP is close to 0.3, the per concept Average Precision (AP) varies a lot. Up to 0.8 or more for "Anchorperson", "Studio_With_Anchorperson" and "Instrumental_Musician", close to 0.1 for many others, and close to 0.01 for "Car_Racing". These differences are partly due to the high and low frequencies of the target concepts tn the test set and to the intrinsic difficulty of detecting them.

In comparison, figure 8 shows the (inferred) con-

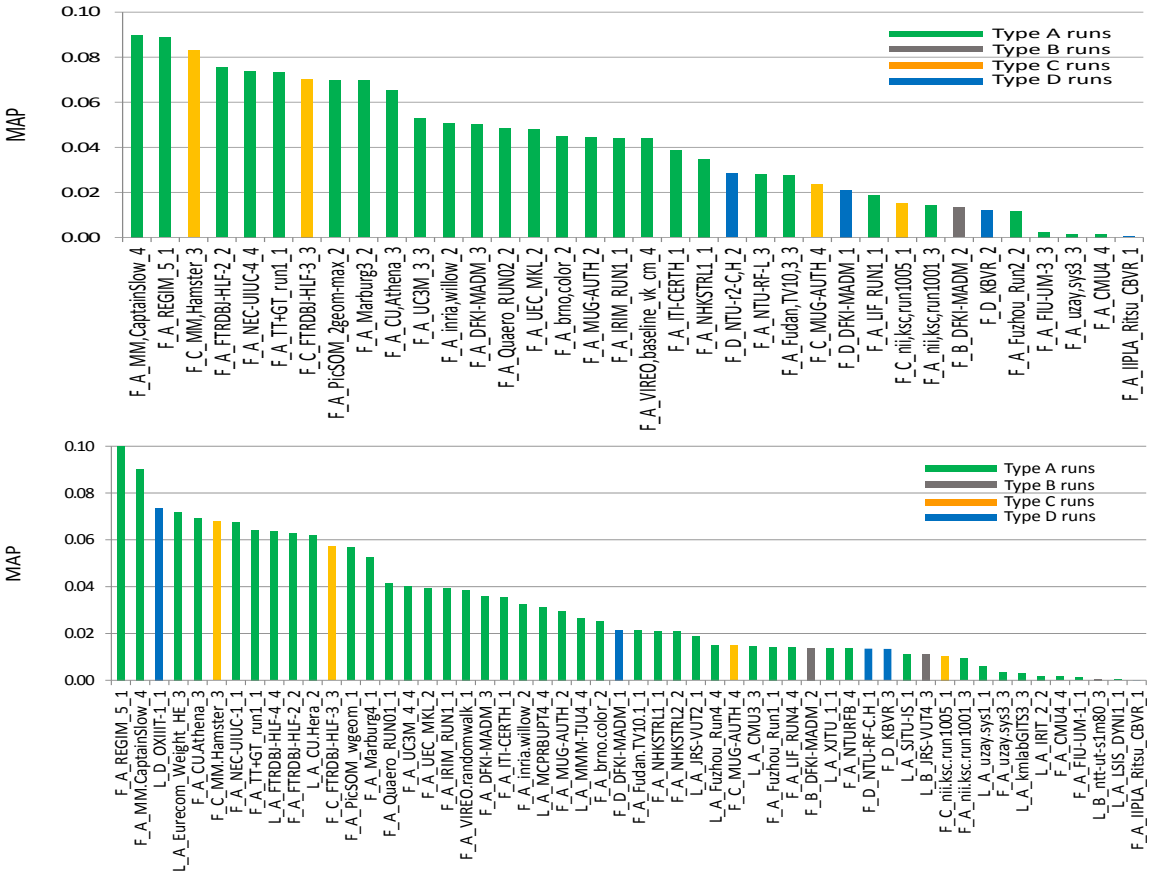


Fig. 1 2010 full (top) and light (bottom) tasks results

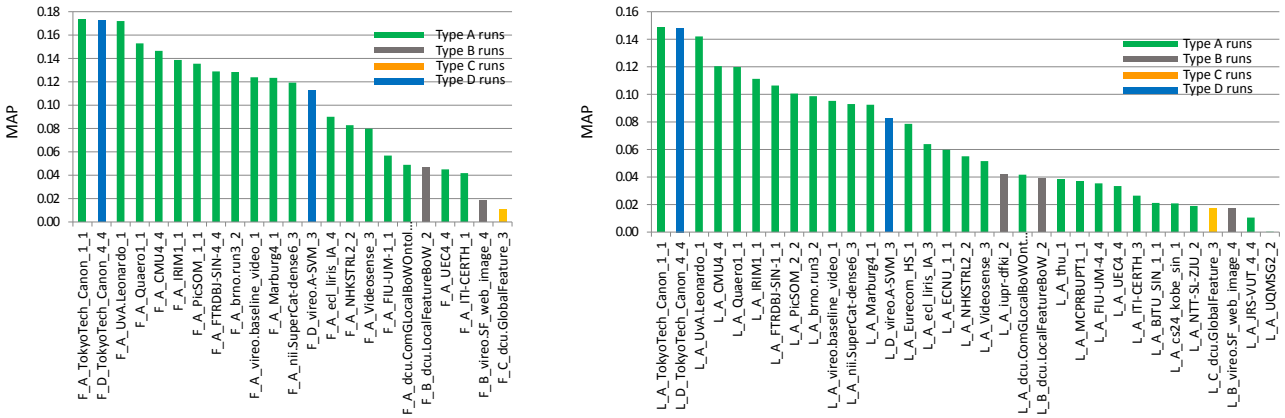


Fig. 2 2011 full (left) and light (right) tasks results

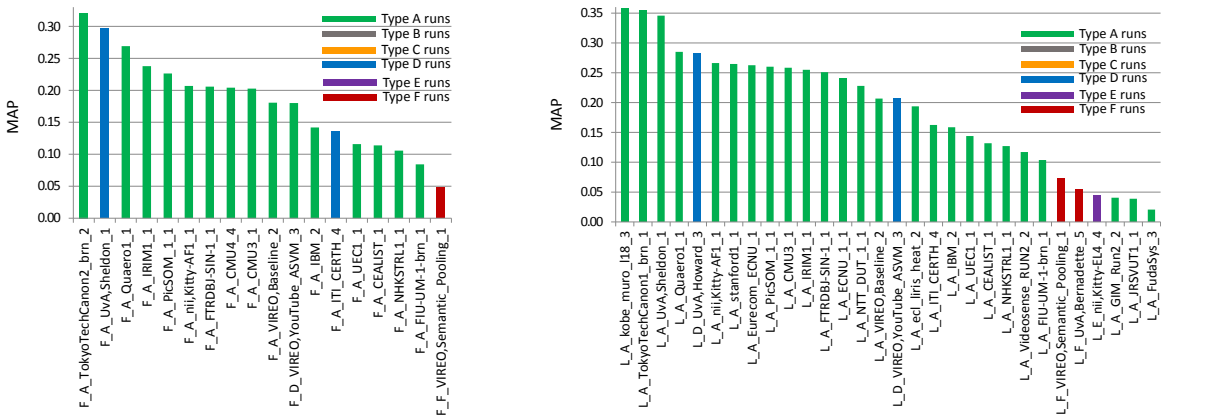


Fig. 3 2012 full (left) and light (right) tasks results

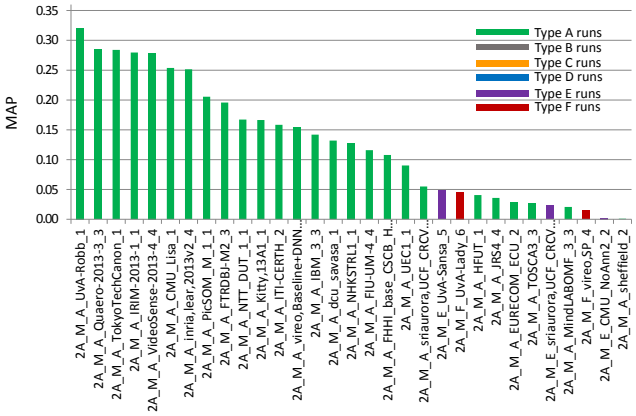


Fig. 4 2013 main task results

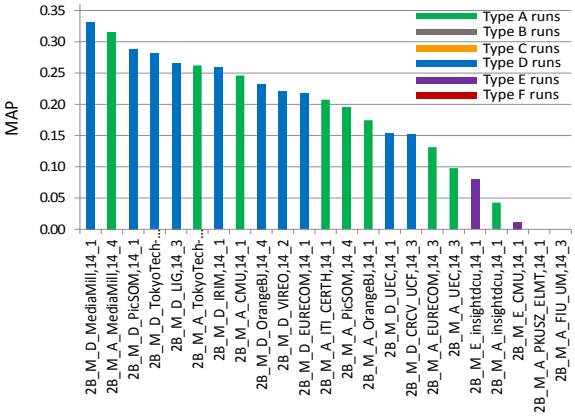


Fig. 5 2014 main task results

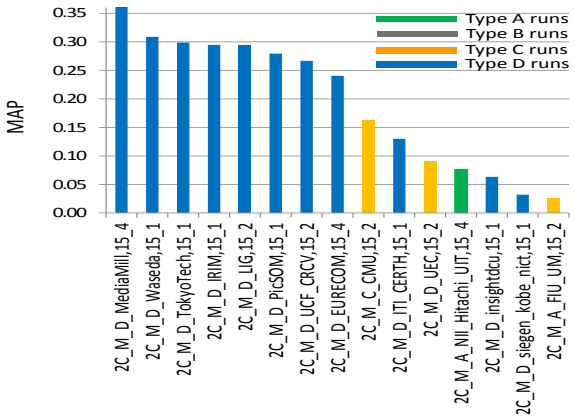


Fig. 6 2015 main task results

cept frequencies in the test collection. These frequencies correspond to the AP of a system making random prediction. Most concept frequencies are below 1% and even below 0.5%. The average concept frequency is of 0.62% while the MAP of the best and median systems are respectively of 36.2% and 24.0%. It can be observed too that concepts with similar frequencies may obtain quite different Average Precisions and vice versa. For instance, "Computers" and "Old_People" have similar frequencies but the AP is much higher for "Computers" indicating that "Old_People" is harder to detect. Similarly, "Instrumental_Musician" and "Studio_With_Anchorperson" have similar Average Precisions but "Studio_With_Anchorperson" is much less frequent indicating that "Instrumental_Musician" is harder to detect. This can be understood by the fact that "Instrumental_Musician" is a true multi-modal target where it is required that the musician can be simultaneously seen and heard. "Basketball" is quite well detected, with an AP of 15.4%, even though it is very infrequent with a frequency of 0.013%.

Figure 7 shows only the results for the top 10 submissions of all participants. Though these include several runs from same participants, they gather results from several different participants and are quite often quite grouped, indicating that the best participants or systems always obtain very similar performances for the same target concepts, even though the median run (at a depth of 29) is significantly lower. This is particularly true for instances of "Airplane", "Boat_Ship", "Demonstration_Or_Protest", "Office", "Hills" or "Quadruped". For some other concepts, the AP varies much more within the top 10. This is the case for: "Cheering", "Government_Leaders", "Motorcycle", "Telephones", "Throwing" or "Flags".

4.3 Approaches

Though, as previously mentioned, the performance of systems cannot be directly compared across years due to changes in test data, target concepts and the amount of annotation data available, significant progress has been achieved over the six years during which the SIN task was run. This is confirmed for the last three years in the context of the progress secondary task as can be seen in section 8 but it is likely that this was also the case for the previous years. The approaches of the participants significantly evolved over time leading to significant increases in systems' performance. All of them rely on supervised learning using the provided training data or other annotated data or both. Though there

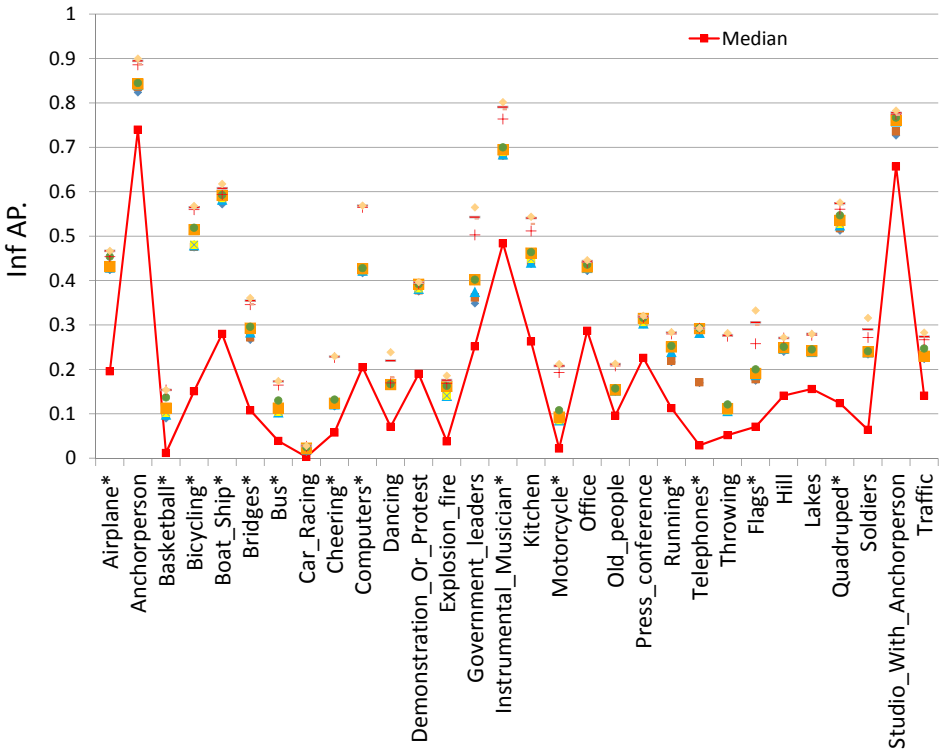


Fig. 7 Top 10 InfAP scores by concept for the 2015 main task. Starred concepts were common between the 2014 and 2015 main tasks.

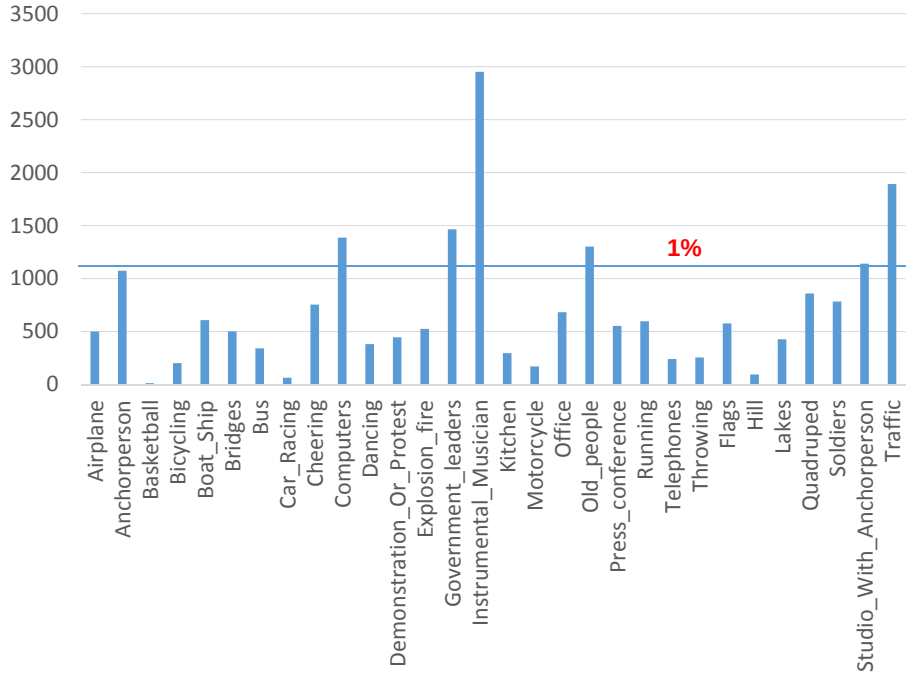


Fig. 8 Inferred concept frequency for the 2015 main task

were lots of variations and particular approaches, three main phases could be observed.

In the first phase, many systems followed the "Bag of Visual Words" approach (BoVW) [13)14)] which consists of applying the "bag of words approach" popular in textual information retrieval. In this approach, local features (or descriptors) are extracted for a number of points or patches in images (key frames) or in video shots, and are aggregated into a single global representation. Local features are first quantized according to a "dictionary" built by clustering features extracted on training data. Images or video shots are then represented as histograms of quantized features. Among the most popular local features are: the Scale Invariant

Feature Transform (SIFT)[15] and its color version[16] for still images, and the Spatio-Temporal Interest Points (STIP)[17] for video shots. These representations can be obtained from sparse sets of points or regions selected using for instance a Harris-Laplace detector or on dense sets following regular grids. Additionally, representations can be computed either on a whole image or separately on various image decompositions including pyramidal ones[16]. Other approaches also involves bag of trajectories.

As alternatives or complements to the BoVW approach, participants used simpler descriptors like color histograms, Gabor transforms or local extraction of semantic information (semantic categories on image patches). A few participants also used audio descriptors, most of which were derived from sequences of Mel Frequency Cepstral Coefficient (MFCC) vectors, either via global statistics (mean, standard deviation, ...) or again via the bag of words approach.

The shot or key frame representations are then used for supervised learning, mostly using Support Vector Machines (SVM) classifiers. Most participants used several different representations (e.g. color, texture, interest points, audio, motion ...) and/or several machine learning methods and fused them for obtaining better results. Fusion methods included early and late fusion[18], and kernel fusion[19], either in flat or in hierarchical ways[20].

In the second phase, following their introduction in still image representation, improved aggregation methods were introduced or designed for video shot representation. These include Fisher Vectors[21], Vectors of Locally Aggregated Descriptors[22], Vectors of Locally Aggregated Tensors[23], and SuperVectors[24]. These methods allowed a significant improvement over the basic BoVW approach, even when using the same local descriptors. These methods rely on the use of GMM representations of the training data which capture more information than the basic BoVW approach.

In the third phase, deep learning methods that made a significant breakthrough in still image categorization at ILSVRC 2012[25] were introduced and led to another significant improvement over the classical feature extraction and learning approaches. In contrast to these classical approaches, Deep Convolutional Neural Networks (DCNN) are end-to-end solutions in which both the feature extraction and the classifier training are performed at once. The first layers extract a type of information which is similar to the features/descriptors extracted in the classical approaches. These, called "deep features", turn out to be significantly more efficient that the classical "engineered" ones, even when used with classical machine learning for classifier training. The last DCNN layers performs the final classification with a single network for all the target concepts. The global training of DCNNs guarantees and optimal complementarity between the feature extraction part and the classification part.

The TRECVid training data from the collaborative annotation does not contain enough data for a complete training of a large scale Deep Convolutional Neural Network (DCNN). When tried, this approach performed significantly less well than the two main alternative approaches also used in other domains. The first one consists of partially retraining a DCNN already trained on ImageNet data for adapting it to TRECVid (IACC) data. In this approach, the first layers, corresponding to the feature extraction part, are frozen and only the few last layers are retrained. This is because the deep features trained on ImageNet are very general and does not depend much upon the training data or upon the target concepts while the last layers are much more specific to the set of target concepts. It has been experimentally observed that retraining only very few of the last layers is the best choice, the optimal number being typically only two or even one depending upon the DCNN architecture. The second main alternative to a full DCNN retraining consist in extracting the output of the few last layers and using them just as ordinary features in a classical machine learning (e.g. SVM) approach. Once again, it has been observed that the last two hidden layers and even the final output layer are the best candidates.

Fusion proved to be very efficient when used in conjunction with the deep learning approach. Such fusion can be done in many different ways: late fusion of the different network architectures, late fusion of a same architecture but with different training conditions, late fusion of partially retrained DCNNS and classical classifiers using deep features, late or early fusion of deep features combined with classical classifiers, late fusion of DCNN-based classifiers and fully classical systems using engineered features. Though all of these solutions may have different performances, their fusion almost always outperform the best elementary component with the general rule that the more elements are integrated

in a system, the best performances this system reaches, possibly leading to very high system complexity as this was the case already with the classical approaches.

Other completely independent methods have also been used for further improving the system performance, some of them not really new. Among them: the use of multiple key frames for increasing the chance of identifying the target concept in a video shot and the use of the detection of a concept in adjacent shots for exploiting the local semantic coherency in video contents. In the context of DCNNs, the data augmentation approach has also been used, also leading to a significant performance improvement.

The use of audio and motion (STIP or trajectory-based) features does help in the classical approach but with generally a modest contribution. No use of audio or motion were considered yet in the best performing deep learning based approaches. Ontology relations (implies and excludes) were provided but they did not seem to be used directly by the participants, probably due to the difficulty of integrating hard rules with detection scores. However, these were used in the collaborative annotation for generating 28,864,844 total annotations from the 8,158,517 direct ones. So these relations were used indirectly in the training. Implicit or statistical relations between concepts were also used by some participants.

5. Concept Pair Task

For the 2012 and 2013 edition of the TRECVid benchmark, a secondary concept pair task was offered to SIN participants. This section motivates the task, summarizes results, and highlights the approaches.

5. 1 Motivation

An important motivation for the regular SIN task is to provide semantic tags for video retrieval technologies like filtering, categorization, browsing, and search. While a single concept detection result has been proven by many to be a valuable resource in all these contexts, several video retrieval scenarios demand more complex queries that go beyond a single concept. Examples of concept pairs are $Animal + Snow$, $Person + Underwater$ and $Boat/Ship + Bridges$. Rather than combining concept detectors at query time, the concept pair task strives for detecting the simultaneous occurrence of a pair of unrelated concepts in a video, where both concepts have to be observable simultaneously in a shot. The overall goal of the concept pair task is to promote the development of methods for retrieving shots containing a combination of concepts that do better than just combining the output of individual concept detectors.

While it can be foreseen that existing single concept detectors can also be trained using concept pair annotations, the combination of potential concept pairs is massive. Hence, such a pair-annotation approach seems unfeasible in practice and is therefore discouraged. By design, the concept pair task did not provide any pair annotations to participants.

5. 2 Results

The performance metric for this task is the (inferred) MAP exactly as for the main task. The 2012 edition of the concept pair task received a total of twelve submissions from six different teams. The top run achieved a score of 0.076 while the median score was 0.041. In addition, the MediaMill team from the University of Amsterdam provided four baseline runs using their single-concept run as the basis. The runs simply relied on the first concept occurrence only, the second concept occurrence only, the sum of both concept detector scores, and the product of both concept detector scores. The baseline recognizing pairs by focusing on the first concept only proved to be a surprisingly valuable tactic, ranking third with a score of 0.056. For the pair $Driver + Female\ Human\ Face$ the baseline even came out best. Motivated by the fact that systems for pair detection have difficulty in finding evidence for concept co-occurrence it was decided to continue the secondary task in 2013.

In 2013 participation grew to ten teams, submitting a total of 20 runs. Each participant was requested to submit a baseline run which just combines for each pair the output of the groups two independent single-concept detectors. In addition, the option to indicate the temporal order in which the two concepts occurred in a video shot was offered, but no teams participated in that. The top run in 2013 achieved a score of 0.162. While this seems much better than the score obtained in 2012, it should be noted that the pairs changed and some may have been easier, or less rare, than the ones in 2012. The best performer for the pair $Government\ Leader + Flags$, for example, scored 0.658. Among the teams who submitted baselines, we found that three of them had baselines that achieved better scores than their regular runs, while only two teams had all their regular runs improve over the baseline. The best run simply combined individual concept detector scores by their product. As their was no experimental evidence after two editions

of the task that dedicated approaches could outperform the simple baselines it was decided to stop the concept pair task after the 2013 edition for the time being.

5.3 Approaches

The majority of runs in the concept pair task focused on combining multiple individual detectors by well known fusion schemes, including sum, product and geometric mean. Some considered compensation for quality and imbalance in training examples of individual detectors by weighted fusion variants. Other approaches learned the pair directly from the intersection of annotations for the individual concepts or gathered examples containing the pair from the web. Among the more unique approaches was the submission from CMU which considered looking at many concepts, beyond just the pair, to enhance the prediction of pair-concepts using several semantically related concepts. Also unique was the submission by the MediaMill team, which tried to reduce the influence of the global image appearance on individual concept detectors, by considering spatio-temporal dependencies between localized objects. Unfortunately, none of these approaches were able to outperform the simple combination baselines. Time will tell whether a concept pair is more than the sum of its parts.

6. Concept Localization Task

In order to encourage more precise concept detectors, in 2013 a new secondary task was initiated for localizing the occurrence of visual concepts in both the temporal and spatial domains. The main goals of this secondary task are to test the precision of concept detectors to the frame (temporal) and bounding box (spatial) levels instead of just the shot-level, as in the main SIN task. The better the systems do their design of precise detectors, the more re-usable they are as they become less dependent on the video context. During 2013 and 2014 this secondary task was run where systems participating in the SIN task had the option to submit runs to localize the first 1 000 shots. In 2015 the organizers decided to run this as an independent secondary task where systems were given a set of relevant shots and asked to return localization result sets. In total 10 concepts were chosen for localization. In the following sections we discuss in more details the task, data, evaluation framework, metrics and results of participating teams from 2013 to 2015.

6.1 Task definition

This secondary task can be described as follows: for each visual concept from the list of 10 designated for localization, and for each I-Frame within the shot that contains the target, return the x,y coordinates of the upper left and lower right vertices of a bounding rectangle which contains all of the target concept and as little else as possible. Systems may find more than one instance of a concept per I-Frame and then may include more than one bounding box for that I-Frame, but only one was used in the judging since the ground truth contained only 1 per judged I-Frame, the one chosen by the NIST assessor that was supposed to be the most prominent (e.g., largest, clearest, most central, etc.). Assessors were asked to stick with this choice if a group of targets were repeated over multiple frames unless the prominence changes and they have to change their choice.

6.2 Data

For this secondary task we used the same test data sets (IACC.2.A, IACC.2.B, IACC.2.C) as used for SIN from 2013-2015 as the basis for the localization task.

6.3 Evaluation framework

Figures 9 and 10 show the evaluation framework at NIST for the localization secondary task in 2013, 2014 and 2015 respectively. In 2013 for each shot found to contain a localization concept in the main SIN task, a sequential percentage (22 %) subset of the I-Frames beginning at a randomly selected point within the shot was selected and presented to an assessor. However, in 2014 and 2015, a systematic sampling was employed to select I-frames at regular intervals (every 3rd I-frame in 2014 and every alternate I-frame in 2015) from the shot.

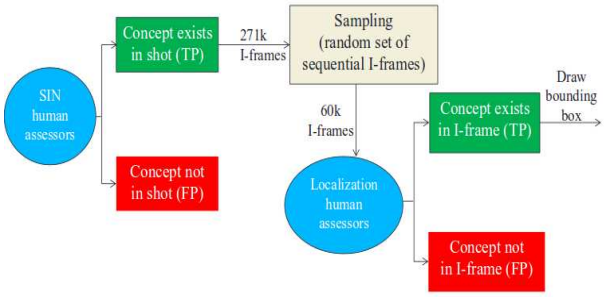


Fig. 9 2013-2014 Evaluation Framework

For each image the assessor was asked to decide first if the frame contained the concept or not, and if so, to draw a rectangle on the image such that all of the visible concept was included and as little else as possible.

In accordance with the secondary task guidelines, if

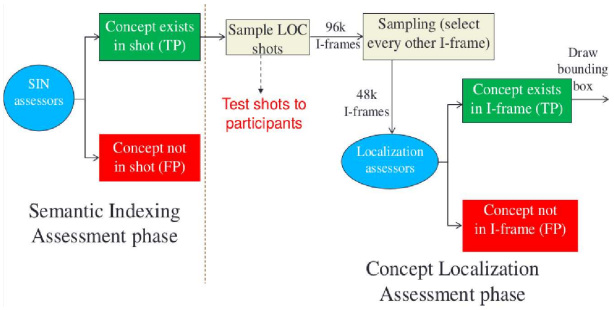


Fig. 10 2015 Evaluation Framework

more than one instance of the concept appeared in the image, the assessor was told to pick just the most prominent one and to box it in. Assessors were told that in the case of occluded concepts, they should include invisible but implied parts only as a side effect of boxing all the visible parts.

Early in the assessment process it became clear that some additional guidelines were needed. For example, sometimes in a series of sequential images the assessor might know from context that a blurred area was in fact the concept. In this case we instructed the assessor to judge such an image as containing the concept and to box in the blurry area.

A minimum of 5 assessor half-days for each of the 10 concepts to be judged was planned (total of 200 labor hours). This was based on some preliminary tests at NIST where it was estimated that each assessor could judge roughly 6,000 images in the time allotted.

Table 3 describes, for each concept, the total number of shots judged to contain the concept and the number of I-Frames comprised by those shots from 2013-2015. Note that the two concepts "Chair" and "Hand" were replaced in 2015 by "Anchorperson" and "Computer" due to the very high frequency of occurrence of "Chair" in the test collection and the ambiguity of the definition of the concept "Hand" (A close-up view of one or more human hands, where the hand is the primary focus of the shot).

Table 3 Number of TP shots and I-frames per concept

Name	*True shots*	*I-Frames*
Airplane	594	10,229
Boat_Ship	1,296	2,917
Bridges	662	884
Bus	561	12,027
Chair	2,375	93,206
Hand...	1,718	20,266
Motorcycle	584	12,086
Telephones	508	19,163
Flags	1,219	41,886
Quadruped	1,233	50,448
Anchorperson	300	14,119
Computers	300	15,814

6.4 Measures Used

Temporal and spatial localization were evaluated using precision, recall and f-score based on the judged I-frames. The I-frame is judged as a true frame temporally if the assessor can see the concept. The spatial recall and precision is calculated using the overlap area between the submitted bounding box and the ground truth box drawn by the assessor. NIST then calculated

an average for each of these score values for each concept and for each run.

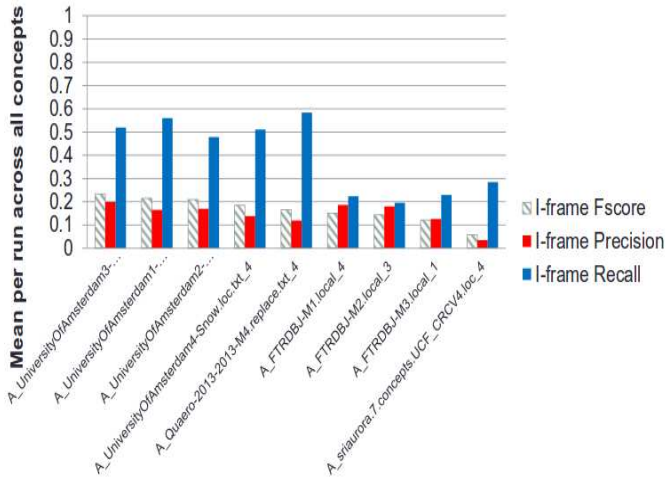
6.5 Evaluations and Results

In this section we summarize the participants' results from 2013-2015 by the type of localization measured. In general, 4 teams finished the first year localization secondary task, submitting total of 9 runs while 1 team finished the second year secondary task with 4 submitted runs. In 2015 when the task became independent from the semantic indexing task, 6 teams finished with a total of 21 runs. As the results of 2014 may not indicate real conclusions about systems performance because only 1 team finished the task we will skip the results from that year and discuss only the 2013 and 2015 results.

(1) Temporal Localization Results

Figures 11 and 12 show the mean precision, recall and F-scores of the returned I-frames by all runs across all 10 concepts in 2013 and 2015 respectively.

Fig. 11 2013: Temporal localization results by run



In 2013 all runs reported much higher recall (reaching a maximum above 50%) than precision or F-score except 1 team (FTRDBJ) which had close scores for the 3 measures. Lower precision scores (maximum 20%) indicate that most runs returned a lot of non-relevant I-frames that did not contain the concept. In 2015 systems reported much higher F-score values compared to the previous two years as 9 out of 21 runs scored above 0.7, and 8 runs scored above 0.6 F-score. We believe these high scores are side-effect of only localizing true positive shots (output of the semantic indexing task) compared to localizing just raw shots (as in 2013-2014) which may include true positive as well as true negative

Fig. 12 2015: Temporal localization results by run

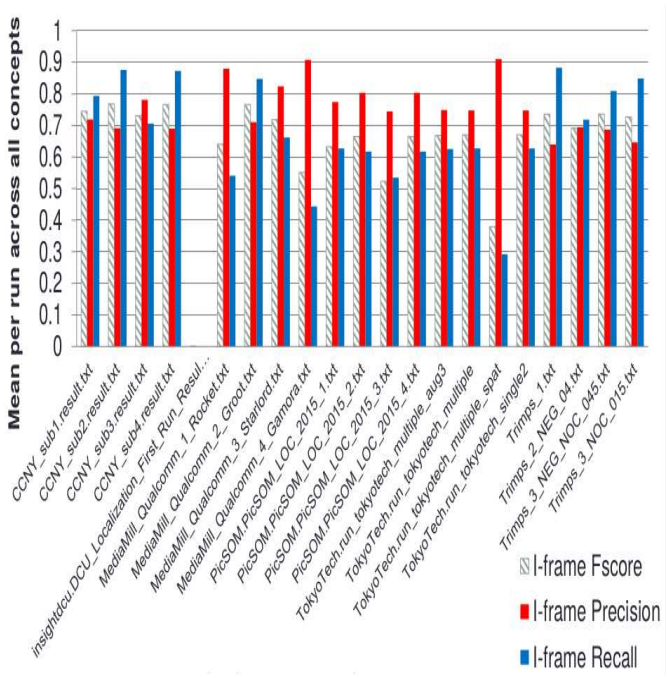
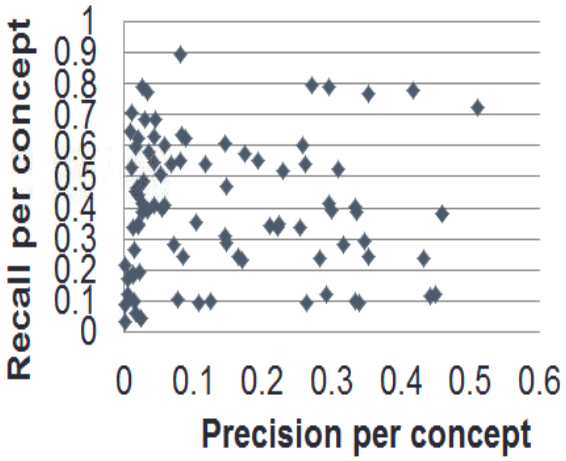


Fig. 13 2013: temporal precision and recall per concept for all teams



concepts.

To visualize the distribution of recall vs precision, we plotted the results of recall and precision for each submitted concept and run in Figures 13 and 14 for 2013 and 2015 respectively. We can see in Figure 13 that the majority of systems submitted many non-target I-frames, achieving high recall and low precision while very few found a balance. However, in 2015 most concepts achieved very high values for both precision and recall (above 0.5).

Fig. 14 2015: temporal precision and recall per concept for all teams

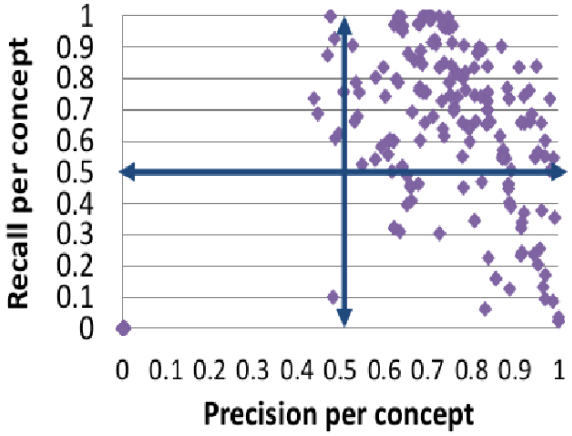
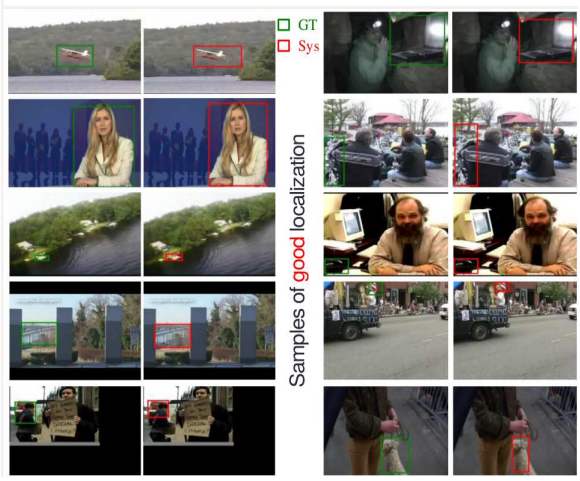


Fig. 15 Visual samples of good results



(2) Spatial Localization Results

Figures 15 and 16 show sample results of good and less good spatial localization results, respectively. These sample results are shown for the 10 concepts listed in table 3, "chair" and "hand" being excluded. We tried to pick hard positive examples in Fig 15

Fig. 16 Visual samples of less good results

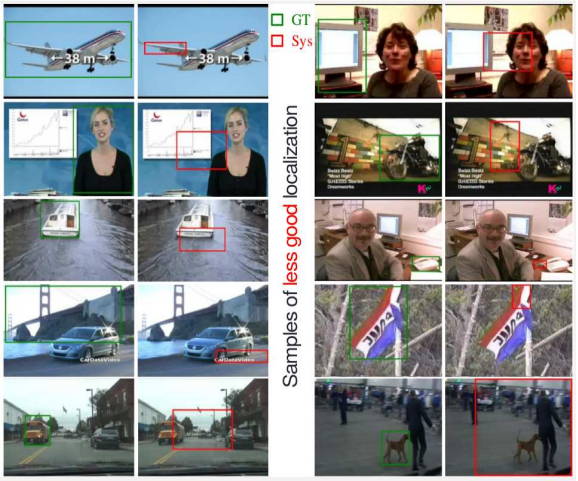
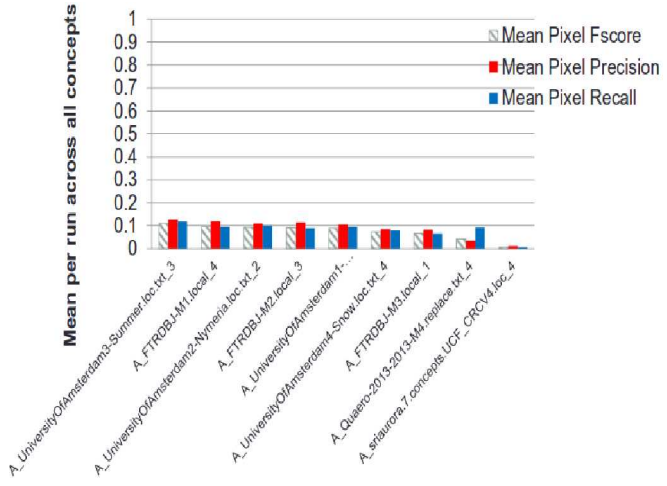
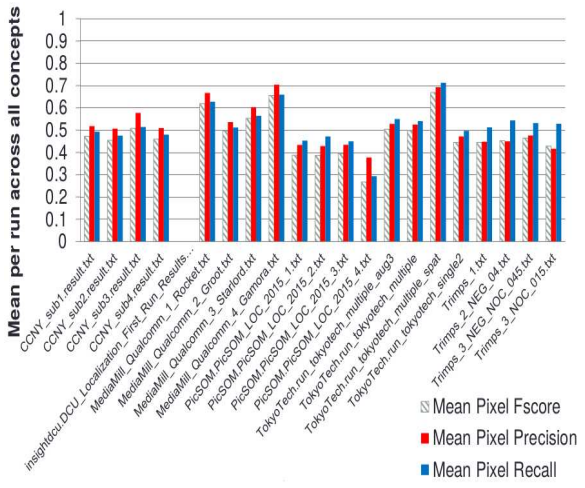


Fig. 17 2013: Spatial localization results by run



(small size, occluded,low illumination, etc) to demonstrate how a sophisticated localization system can perform while we picked easy examples in Fig 16 (centered, big, clear, etc) where results are not good. This variation in performance shows the gap between a top system and low ranked system. Figures 17 and 18 show the performance by run for spatial localization (correctly returning a bounding box around the concept). In 2013 scores were much lower than for the temporal measures and barely reaching above 10% precision. This indicates that finding the best bounding box was a much harder problem than just returning a correct I-frame. In 2015 the F-scores range was less than the temporal F-score range but still higher than the previous two years. Overall, 8 out of the 21 runs scored above 50% and another 8 runs exceeded 40%. The distribution of recall vs precision performance in figures 19 and 20 shows an interesting observation that systems are good at submitting an accurate approximate bounding box size which overlaps with the ground truth bounding box coordinates. This is indicated by the cloud of points in the direction of positive correlation between precision and recall. It can also be shown that in 2015, performance is much better as the distribution of points are moving away from low precision and recall values (less than 0.2) which is on the contrary obvious in 2013.

Fig. 18 2015: Spatial localization results by run



6. 6 Approaches

Most approaches by participating teams started by applying selective search[26] or EdgeBox[27] algorithms to extract a set of candidate boxes independent from the concept category. Features are then extracted from proposed boxes either in a bag of words framework or more recently using deep learning models such as VGG-16,

Fig. 19 2013: spatial precision and recall per concept for all teams

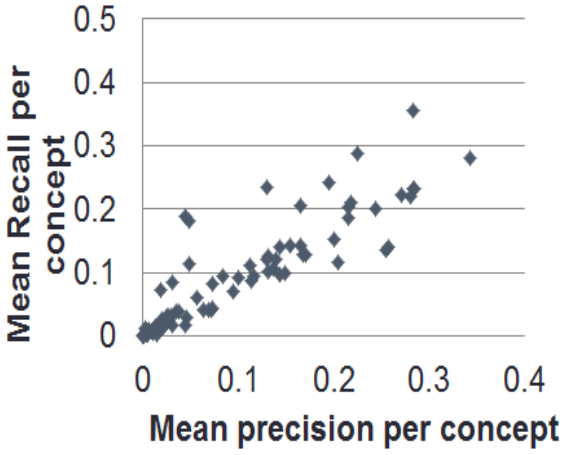
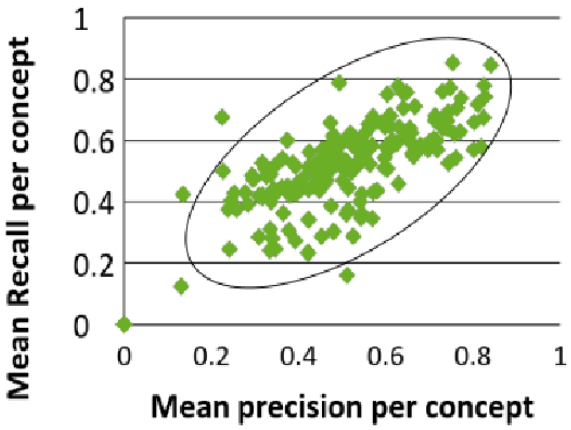


Fig. 20 2015: spatial precision and recall per concept for all teams



fast-RCNN (Region-based Convolutional Neural Networks) or Inception deep neural networks[28]. Support vector machines are usually applied as a final layer for classification. In addition, few teams employed Deformable Part-based models[29] with color or texture features. Deep learning-based approaches, especially the RCNN-based ones, performed the best.

6.7 Summary and Observations on Localization Tasks

The localization secondary task was a successful addition to the semantic indexing main task in 2013 and 2014 and it was decided to run it independently in 2015. In general, detecting the correct I-frames (temporal) only was easier than finding the correct bounding box around the concepts in the I-frames (spatial) and overall, systems can find a good approximate bounding box size that overlaps with the ground truth box but still not with high precision.

In 2015 the scores were significantly higher, mainly because we aimed to make systems just focus on the localization task, bypassing any prediction steps to decide if a video shot included the concept or not as was done in the previous two years in the main semantic indexing task. This may have caused the task to be relatively easy compared to a real-world use case where a localization system would have no way to know beforehand if the video shot already included the concept or not. In future localization tasks we plan to give systems raw shots (which may include true positive or true negative concepts) simulating a semantic indexing predicted shot list for a given concept. We also plan to test systems on a new set of concepts which may include some actions which span much more frames temporally compared to only objects that may not include much motion.

7. No Annotation Task

For the 2012 to 2015 issues of TRECVid, a "no annotation" secondary task was offered to SIN participants. This section describes how that task worked and the outcomes.

7.1 Motivation

The motivation behind launching a "no annotation" secondary task is a reflection of the difficulty associated with finding good training data for the supervised learning tools which are used in automatic concept detection. As seen throughout this paper, and especially in subsection 2.6, the overhead behind manual annotation of positive, and even negative, examples of concept occurrence is huge. The potential for automatically harvesting training data for supervised learning from web resources has been recognised by many, including the first such work by[30],[31],[32] and subsequently by others.

With this in mind, TRECVid offered a secondary SIN task in which no training data for the concepts was provided to participants. There were two variations of the task offered in each of the 4 years, described earlier in section 4 and repeated here:

E used only training data collected automatically using only the concepts' name and definition;

F used only training data collected automatically using a query built manually from the concepts' name and definition.

What is intended here is that participants are encouraged to automatically collect whatever training data they can, and most will use web resources like word-based image search or word-based search of video resources such as YouTube. This proposition is attractive because it means that in theory there is effectively no restriction to the range, or the type of semantic concepts for which we can build detectors for video and this opens up huge possibilities for video search.

The potential downside to this idea is that the efficacy of these detectors will depend on how accurately participants could locate a good quality set of training data. With manual annotation of training data we expect the annotations to be accurate and there will be few, if any, false positives whereas with automatically-collected training data we are at the mercy of the techniques that participants use to harvest such data. In particular, for abstract concepts this will be even more difficult and even for semantic concepts which refer to (physical) objects like "motor car", "tree" or "computer screen", it is a challenge to automatically locate many hundreds of positive examples with no false positives creeping into the training set. However with the quality of image search on major search engines improving constantly, some of that earlier work in the area like that reported in[30],[31],[32] is already quite dated in that they were then dealing with a level of image search quality which is now much improved. An additional problem to the level of noise in automatically crawled data is the possible domain mismatch between the general material that can be gathered from the web and the specific domain for which we may want to build a concept detector for.

7.2 Results

Table 4 shows the number of runs submitted by par-

ticipants for the E- and F-type SIN task condition, for each of the 4 years this secondary task was offered.

Table 4 Number of runs submitted in the "no annotation" secondary task

Year	*Type E*	*Type F*
2012	1	4
2013	6	3
2014	4	0
2015	0	0

From this we can see we had very low participation with only 18 runs from just a few participants over 3 of the 4 years this was offered. What was interesting about those results was the performance, as measured in terms of mean InfAP. In 2012 the best-performing category A result was 0.32 infAP with a median across submissions of 0.202 while the best-performing category F result was 0.071, with a median of 0.054. The "no annotation" results fall far short of the full category A but for a first running of the secondary task, this was encouraging. By 2014 (there were no results submitted in 2015), the best category E submission scored 0.078 against a best category A submission of 0.34 (mean 0.217). Once again these results are encouraging but with low interest in the task and no participation in its last year, we may have already tapped into all the interest that there might be in this topic.

7. 3 Approaches

For the (limited) number of participants who submitted runs in this task, some used the results of searches to YouTube as a source of training data, others used the results of searches to Google image search, and some used both.

One of the participating teams (the MediaMill group at the University of Amsterdam) investigated three interesting research questions, described at[33]. They found that ...

- Tagged images are a better source of training data than tagged videos for learning video concept detectors;
- Positive examples from automatically selected tagged images shows best performance;
- Negative training examples are best selected with a negative bootstrap of tagged images

One of the things that this secondary task has raised is the question of whether a no annotation approach to determining concept presence or absence is better applied *a priori* at indexing time, as in this task, or dynamically at query time. One of the disadvantages of indexing video by semantic concepts in advance of searching is that we need to know, and define, those concepts and that limits subsequent searching and video navigation to just those concepts that have been built and applied to the video collection. Building concept detectors at query time allows concepts to be dynamically constructed, if this can be achieved with reasonable response time.

Recent work such as the one by[34] has shown that it is possible to take a text query and download several hundreds of top-ranked images from Google image search, compute visual features of those images on-the-fly and use these as positive examples for building a classifier which is then applied to a video collection to detect the presence (and absence) of video shots containing the concept represented by the Google image search query, and to do all this within a couple of seconds while the searcher waits for query results. In the work reported to date this is shown to work well for visually distinct objects like "penguin", "guitar" or "London bus" where the issues of quality of the training set in terms of how many false positives creep into the top-ranked images when searching for penguins, guitars or London buses, is not so important. Further work to refine and improve the training set will mean that more challenging concepts should be detectable and this would offer a real alternative to what was promoted in this secondary SIN task.

8. Progress Task

8. 1 Motivation

Evaluation campaigns like TREC, TRECVid, ImageNet LSVRC and many others are very good for comparing automatic indexing methods at a given time point. The evaluation protocols are usually well designed so that comparisons between methods, systems and/or research teams are as fair as possible. The fairness of the comparison relies for a significant part on the fact that all systems are compared using the same training data (and annotations) and that test data are processed blindly with results being submitted within the same deadline. It also relies on the trust granted to the participants that they do respect the guidelines, especially considering blind processing. While it is acceptable that they have a look at the results for checking that these make sense and for detecting or fixing major bugs, they should never do any system tuning by analyzing them.

This approach implies that when such campaigns are organized periodically, new fresh test data are made

available for each issue because a lot of information can be obtained via the analysis of past results, taking them into account in a new version of the system. Applying the new system on past data will then result in biased and invalid results. This is the approach used for the SIN task (as can be seen in Table 2) and more generally at TRECVid.

While it is good to compare various systems or methods at a given time point, it is also interesting to monitor the overall evolution of the state of the art methods' performance over time. As previously mentioned, it is not possible to do this directly using the results obtained from consecutive issues of TRECVid because they differ on the test samples, on the evaluated categories and/or on the amount of training data. The ideal solution would be that regular participants keep a version of their system from each year and apply it, unchanged, for each of the subsequent years. Even in this case, the comparison would not be meaningful if new training data became available in the intervening period. For practical reasons, it is often complicated to maintain over years, a number of previous versions of the systems and, in the best cases, some participants are able to make one reference submission using their best system from the previous year. Some studies have shown that significant progress has been achieved over time in the past[35]. However, these have been made *a posteriori*, and while their conclusions are valuable, they did not strictly follow the blind submission process. Also, they concerned only submissions from a single participant.

The "progress" secondary task was developed following the feedback from a number of participants from 2010 to 2012. Its goal was to obtain meaningful comparisons between successive versions of systems and to accurately measure the performance progress over time. It was conducted on the 2013 to 2015 issues by:

- releasing the test data for the three 2013 to 2015 issues at once;
- freezing the training data and annotation sets (no new annotations were made available in 2014 and 2015);
- freezing the concept set for which submissions were requested;
- requiring participants each year to directly submit runs for the current issue and for all the available next issues (i.e. in 2013, participants submitted runs for the 2013, 2014 and 2015 test collections; in 2014, they submitted runs for the 2014 and 2015 test collections; in 2015, they submitted runs only for the 2015 test collection).

Apart from the fact that some submitted runs are anticipated submissions for future years, this secondary task is exactly the same as the main SIN task described in section 4. Submissions to the progress task corresponding to the current year are the same as those for the main task by the same participant. Submissions made by a participant for the future years are included in the pool of submissions for these future years. These anticipated submissions have been filtered out in the presentation of the results in section 4 but they were included in the same evaluation process, including their insertion in the pooling process for assessment.

Submitting to the progress secondary task required little effort from participants just running their systems on one, two or three slices of the test data instead of only just one, while the main work was in the design, the training and the tuning of their systems. The rule of not using new annotations for the 2014 and 2015 submissions was specific to the progress task. Some participants to the main task that did not submit to the progress task and actually used the 2013 and/or 2014 assessment as additional annotations, especially for parameter tuning by cross-validation on them. This possibly induced a small disadvantage for the participants to the main task that strictly followed the progress task protocol.

8.2 Results

Six groups participated in the progress task by submitting anticipated runs in 2013 and 2014: Eurecom, IRIM, ITI_CERTH, LIG/Quaero, UEC and insightdcu. Figure 21 shows the performance obtained on the 2015 test collection with their 2013, 2014 and 2015 systems. For most of them, a significant performance improvement is observed. Some of the points, e.g. Eurecom and UEC 2013 submissions, are "outliers", their low performance being due to bugs in their submissions. In the case of Eurecom, IRIM and LIG-Quaero, most of the performance gains come from the use of more and more deep features. For IRIM and LIG-Quaero, it also comes from the use of multiple key frames in 2015. The typical performance gain between 2013 and 2015 is of about 30% in relative MAP value. It was mostly due to the use of deep learning, either directly via partial retraining, or indirectly via the use of deep features, or via combinations of both.

In addition to the official progress task, some participants like the University of Amsterdam often submitted

one run for the current year using their previous year's best system as a baseline. Though this approach does not strictly follow the progress task protocol, it still produces meaningful results that also demonstrate significant progress over years. Additionally, some participants like the University of Helsinki compared the year-on-year progression of their PICSOM system over 10 years[36], including most of the current semantic indexing task period but also the previous High-Level Feature (HLF) detection tasks of TRECVid 2005 to 2009.

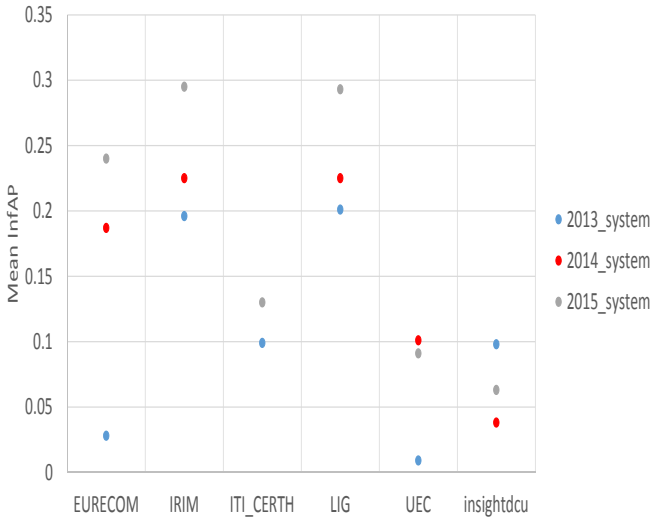


Fig. 21 Progress task results: performance on 2015 test data from 2013, 2014 and 2015 systems.

9. Post-campaign experiments

The TRECVid advisory committee has decided to stop (or suspend) the Semantic Indexing task in 2016. The main reason is that a lot has been learned on this problem for which many techniques are now mature and effective and it is time to move back to the previously suspended main video search task. In the context of the new Ad hoc Video Search (AVS) task, semantic indexing is likely to play a significant role and it will still be indirectly evaluated as a key component.

The data, annotations, metrics, assessment and protocol of the task will remain available for the past TRECVid participants or for new groups that would like to use them for post-campaign experiments. This is similar to what is proposed for the Pascal VOC Challenge[3] closed in 2012 and for which it is still possible to evaluate submissions for the past campaigns and for which an evaluation server is still running and a leaderboard is permanently maintained. This will be slightly different in the case of the TRECVid semantic task. First, the "ground truth" on the test data has been released for the SIN task while it is maintained hidden for Pascal VOC. However, in both cases, the validity of the results rely on the trust granted to the participants that they will not tune their system on the test data; this is a bit harder in the case of VOC but still possible since a number of test submissions are possible. In the case of the TRECVid SIN task, participants to post-campaign experiments should not in any case tune their systems on the test data; for the results to be valid and fair, system tuning should be done only by cross-validation within the development data. A second difference is that there will be no evaluation server; the evaluation will have to be made directly by the participants using the provided ground truth and the `sample_eval` tool available on the TRECVid server.

10. Conclusion

The Semantic INdexing (SIN) task has been running at TRECVid from 2010 to 2015 inclusive with the support of NIST and the Quaero project. It followed the previously proposed High-Level Feature (HLF) detection task which ran from 2002 to 2009[2]. It attracted over 40 participants during the period. The number of participants gradually decreased during the period, while it increased during the previous HLF task, but still 15 groups finished during the last two editions.

The task was conducted using a total of 1,400 hours of video data drawn from the IACC collection gathered by NIST. 200 hours of new test data was made available each year plus 200 more as development data in 2010. The number of target concepts started from 130 in 2010 and was extended to 346 in 2011. Both the increase in the volume of video data and in the number of target concepts favored the development of generic and scalable methods. A very large number of annotations was produced by the participants and by the Quaero project on a total of 800 hours of development data.

In addition to the main semantic indexing task, four secondary task were proposed: the "localization" task, the "concept pair" task, the "no annotation" task, and the "progress" task.

Significant progress was accomplished during the period as this was accurately measured in the context of the progress task but also from the participants' contrast experiments. Two major changes in the methods were observed: a first one by moving from the basic "bag of visual words" approach to more elaborate aggregation methods like Fisher Vectors or SuperVectors,

and a second one with the massive introduction of deep learning, either via partially retrained network or via the use of features extracted using previously trained deep networks. These methods were also combined with many other like fusion of features or of classifiers, use of multiple frame per shot, use of semantic temporal consistency, and use of audio and motion features. Most of this progression was directly made possible via the development data, the annotations, and the evaluations proposed in the context of the TRECVid semantic indexing task.

Acknowledgments

This work was also partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation. The authors wish to thank Paul Over from NIST, now retired, for his work in setting up the TRECVid benchmark and for his help in managing the semantic indexing task.

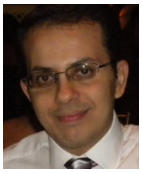
Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

References

1) Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
2) Alan F. Smeaton, Paul Over, and Wessel Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Ajay Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.
3) M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.
4) Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.
5) Stéphane Ayache, Jérôme Gensel, and Georges Quénot. CLIPS-LSR Experiments at TRECVID 2006. In *TREC Workshop on Video Retrieval Evaluation*, Gaithersburg, United States, 2006.
6) Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
7) Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University ADVENT #223-2008-1, August 2008.
8) M. Naphade, J. R. Smith, J. Tesic, Shih F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13:86–91, 2006.
9) Séphane Ayache and Georges Quénot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, mar 2008.
10) Bahjat Safadi, Stéphane Ayache, and Georges Quénot. Active Cleaning for Video Corpus Annotation. In *MMM 2012 - International MultiMedia Modeling Conference*, pages 518–528, Klagenfurt, Austria, January 2012.
11) Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 102–111, 2006.
12) Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, 2008.
13) Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–, Washington, DC, USA, 2003. IEEE Computer Society.
14) G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
15) David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.
16) K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
17) Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September 2005.
18) Cees G.M. Snoek, Marcel Worring, and Arnold W.M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of ACM Multimedia*, Nov. 2005.
19) Stéphane Ayache, Georges Quénot, and Jérôme Gensel. *Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings*, chapter Classifier Fusion for SVM-Based Multimedia Semantic Indexing, pages 494–504. 2007.
20) Sabin Tiberius Strat, Alexandre Benoit, Patrick Lambert, Herv Bredin, and Georges Qunot. Hierarchical late fusion for concept detection in videos. In Bogdan Ionescu, Jenny Benois-Pineau, Tomas Piatrik, and Georges Qunot, editors, *Fusion in Computer Vision*, Advances in Computer Vision and Pattern Recognition, pages 53–77. Springer International Publishing, 2014.
21) Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.
22) Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *23rd IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10)*, pages 3304–3311, San Francisco, United States, 2010. IEEE Computer Society.
23) David Picard and Philippe-Henri Gosselin. Efficient image signatures and similarities using tensor products of local descriptors. *Computer Vision and Image Understanding*, 117(6):680–687, March 2013.
24) N. Inoue and K. Shinoda. A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors. *Multimedia, IEEE Transactions on*, 14(4):1196–1205, Aug 2012.
25) Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
26) J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.
27) C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014.
28) Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

29) Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.
30) Adrian Ulges, Christian Schulze, Daniel Keysers, and Thomas M. Breuel. *Computer Vision Systems: 6th International Conference, ICVS 2008 Santorini, Greece, May 12-15, 2008 Proceedings*, chapter A System That Learns to Tag Videos by Watching Youtube, pages 415–424. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
31) Arjan T Setz and Cees GM Snoek. Can social tagged images aid concept-based video search? In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1460–1463. IEEE, 2009.
32) Jianping Fan, Yi Shen, Ning Zhou, and Yuli Gao. Harvesting large-scale weakly-tagged image databases from the web. In *CVPR*, volume 10, pages 802–809, 2010.
33) Svetlana Kordumova, Xirong Li, and Cees G. M. Snoek. Best practices for learning video concept detectors from social media examples. *Multimedia Tools and Applications*, 74(4):1291–1315, 2014.
34) Ken Chatfield, Relja Arandjelović, Omkar Parkhi, and Andrew Zisserman. On-the-fly learning for visual search of large-scale image and video datasets. *International journal of multimedia information retrieval*, 4(2):75–93, 2015.
35) Cees G. M. Snoek and Arnold W. M. Smeulders. Visual-Concept Search Solved ? *IEEE Computer*, 43(6):76–78, June 2010.
36) Ville Viitaniemi, Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. Advances in visual concept detection: Ten years of trecvid. *Advances in Independent Component Analysis and Learning Machines*, page 249, 2015.



George Awad George Awad is a computer scientist at Dakota Consulting, Inc and contractor for the National Institute of Standards and Technology. He is the current TRECVID project leader and have been supporting the TRECVID project since 2007 as Guest Researcher. He has Msc. in Computer Engineering (2000) from AASTMT (Arab Academy for Science, Technology & Maritime Transport) and PhD in Computer Science (2007) from Dublin City University. His past research interests included image compression, gesture & sign language recognition and hand and face tracking. Currently his main research activities includes evaluating video search engines on different real-world use-case scenarios and using real-world data sets.

Cees G. M. Snoek Cees G. M. Snoek is a director of QUVA Lab, the joint research lab of the University of Amsterdam and Qualcomm on deep learning and computer vision. He is also a principal engineer at Qualcomm and an associate professor at the University of Amsterdam. His research interests focus on video and image recognition. Dr. Snoek is the lead researcher of the MediaMill Semantic Video Search Engine and recipient of several career awards. Cees is a general chair of ACM Multimedia 2016 in Amsterdam.

Alan F. Smeaton Alan Smeaton is Professor of Computing at Dublin City University where he is also Director of the Insight Centre for Data Analytics. He holds a PhD in Computer Science from University College Dublin (1987) and is an elected member of the Royal Irish Academy. His research interests cover all kinds of information systems which support, in some way, human memory. In 2015 he was the winner of the Royal Irish Academy Gold Medal for Engineering Sciences and in 2015 he was also general chair of ACM Multimedia in Brisbane.

Georges Quénot Georges Quénot is a senior researcher at CNRS (French National Centre for Scientific Research). He has an engineer diploma of the French Polytechnic School (1983) and a PhD in computer science (1988) from the University of Orsay. He currently leads the Multimedia Information Indexing and Retrieval group (MRIM) of the Laboratoire dinformatique de Grenoble (LIG) where he is also responsible for their activities on video indexing and retrieval. His current research activity includes semantic indexing of image and video documents using supervised learning, networks of classifiers and multimodal fusion.