Noname manuscript No.
(will be inserted by the editor)

Instance Search Retrospective with Focus on TRECVID

George Awad · Wessel Kraaij · Paul Over · Shin'ichi Satoh

Received: date / Accepted: date

Abstract This paper presents an overview of the Video Instance Search benchmark which was run over a period of 6 years (2010-2015) as part of the TREC Video Retrieval (TRECVID) workshop series. The main contributions of the paper include i) an examination of the evolving design of the evaluation framework and its components (system tasks, data, measures); ii) an analysis of the influence of topic characteristics (such as rigid/non rigid, planar/non-planar, stationary/mobile on performance; iii) a high-level overview of results and best-performing approaches. The Instance Search (INS) benchmark worked with a variety of large collections of data including Sound & Vision, Flickr, BBC (British Broadcasting Corporation) Rushes for the first 3 pilot years and with the small world of the BBC Eastenders series for the last 3 years.

Keywords instance search · multimedia · evaluation · TRECVID

George Awad
Dakota Consulting, Inc., 1110 Bonifant Street, Suite 310, Silver Spring, MD 20910; National Institute of Standards and Technology
E-mail: gawad@nist.gov

Wessel Kraaij
TNO, The Hague, the Netherlands, Leiden University, the Netherlands E-mail: kraaijw@acm.org

Paul Over
National Institute of Standards and Technology (*Retired*)

Shin'ichi Satoh
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan E-mail: satoh@nii.ac.jp

1 Introduction

Searching for information in digital video has been a challenging research topic since the mid-nineties. Research started both in the domain of searching news video collections [25] as well as in the area of defense and public safety [15]. An important focus in the computer vision community has been on recognizing and tracking moving objects. The declining costs of digitizing video archives, later on the availability of digital video,and more recently high definition (HD) video becoming a commodity for consumers on mobile phones have given rise to a tremendous increase in the amount of digital video.

1.1 TRECVID: Measuring progress of digital video search technology

The importance of standard test collections for measuring progress in performance of digital video search technology was recognized by the TREC (Text REtrieval Conference) community which spawned TRECVID, the leading evaluation benchmark conference on search related problems in digital video. In the early years, search performance was dominated by taking advantage of the textual elements associated to news video, such as open captions, metadata and automatic speech recognition. Transcribing the visual content of a video was still in its infancy. TRECVID fostered the development of generic concept detectors in the high-level feature extraction task, later renamed as semantic indexing task. In this task, the challenge was to recognize objects, such as cars, scenes (outdoors or indoors) and simple activities such as walking / running. Core challenges have been to decide whether a certain video segment (usually a shot)

contains a certain object (car, boat, etc.). So the task is to annotate video segments with class labels. The standard approach to develop such semantic detectors is to start from a substantially large sample of positive examples of the concept, covering the inherent variety of visual appearance. The variety in visual appearance can differ dramatically across 'concepts', e.g., there are many types of boats, so perhaps the best feature to recognize all these different boats is to recognize an object in a water scene. On the other hand the variety in visual appearance of a concept like US flag is much lower. The second step is to extract low level features from the example images and learn a discriminative classifier.

Describing video using learned concept classifiers is a technology that is still under development. After a decade of research and development, the state of the art video indexing systems can now detect several thousands of concepts with a precision that makes them useful for content-based video search. However, it is clear that challenges remain. An important problem is that the performance of concept detectors drops significantly in a video collection that has different characteristics (e.g., genre, production style, etc). In addition, concept detectors still rely for a large part on the most prevalent visual context. This makes it difficult to construct queries that assume compositional semantics such as 'horse AND beach'. Finally, the fact that learning classifiers is computationally intensive, makes the concept detector pipeline technology less attractive for ad-hoc queries for new visual objects where fast retrieval result is crucial (such as searching surveillance video).

1.2 Motivation for the TRECVID "instance search" task

The need for evaluating a technology for fast video search and retrieval of precise visual objects (entities) given a visual example has been recognized by TRECVID, and led to a pilot task "instance search" in 2010. The term "instance search" is not self-explanatory. After all, most video search use cases require finding "instances" of some object, person, or location in video. But the notion of instance search, as used in TRECVID, is distinct in that it limits the search to instances of *one specific object, person, or location.* This contrasts with generic ad hoc search in which any instance of *any member of a class of objects, persons, or locations* will satisfy the search. An instance search might be looking for shots of this particular dog, while a generic ad hoc search looks for shots of any dog.

The core notion of instance search was historically widened to treat different objects, if manufactured to be indistinguishable, as though they were in fact a single object, e.g., logos. As operationalized in TRECVID, the instance search task was also narrowed to assume as a starting point a description of the needed video based primarily on a very small set of image/video examples - no significant textual description of the needed video is included. It is essentially a form of query by visual example.

Purported use cases for instance search include business intelligence [Where do our (competitor's) products logos appear?], exploring personal, public, security, forensic video collections [Where else does this person appear? What other video was taken in this room?], etc.

Although the term "instance search" finds its main use starting in 2010 in connection with the TRECVID Instance Search Task, work on the problem predates this. For example, earlier studies experimented with object and scene retrieval in two feature-length movies [64], with person-spotting and automatic face recognition for film characters [63], [1], with naming characters in TV video [18], with object (landmark) retrieval in an image collection (Flickr) [54].

A number of considerations spurred the inclusion of the instance search task in TRECVID 2010. First of all, TRECVID had put significant focus on closing the semantic gap for video search. Others, such as the PASCAL Visual Object Classes (VOC) evaluation had focused on similar issues for still images. The TRECVID high-level feature extraction task and ad hoc search tasks had seen a steady increase in performance, but were still considered much more difficult than searching text and concept detectors' performance still depended on the specific dataset. In parallel more low-level tasks such as shot boundary detection and content-based copy detection had been evaluated over several years. These tasks were simpler and participants had demonstrated good results; perhaps partly because only lower-level visual analysis was involved, without the need for class-level abstraction.

Since lower-level visual analysis was getting more mature, it seemed interesting to explore how these techniques could be used to support search based on visual examples; the instance search task was an example in this direction. We expected such a task to be easier than the ad hoc search task, but more difficult than e.g. content-based copy detection. [43] had already shown the power of various local descriptor techniques for the comparison of images of 2D "scenes". In the meantime commercial applications of these techniques such as logo recognition in sports TV coverage or the recognition of landmarks, wine labels, books by your mobile phone camera [7] had become available. The fundamental hypothesis for the instance search task was

that local descriptor techniques (and extensions being developed) could be improved/extended to effectively search for instances of a certain type in video footage by giving an example clipping from a still image or from a video clip.

The aim of this paper is to provide a retrospective of the TRECVID 'Instance Search (INS)' task. In this section we have provided the original motivation for the task, and the further development of the task will be described in Sections 3 and 5. In addition, Section 2 provides a concise overview of relevant research in computer vision and multimedia information retrieval, in order to sketch the developments of models and algorithms that are typically used for a search by visual example system. Section 6 provides an overview of the experiments carried out by the various teams in the 2010-2015 TRECVID evaluations, with extra attention for the more successful systems. Finally, the main findings and recommendations are summarized in Section 7.

2 Related Work

2.1 Overview: Image Search by Visual Example

Image search by visual example, also known as content-based image retrieval (CBIR), has been intensively studied for decades. The basic idea of CBIR is to search images in archives having sufficiently high visual similarities to visual examples, i.e., query images. Users may expect to retrieve images which are semantically similar to visual examples. However, due to the variation in appearance (e.g., a chair can have many forms) the so-called semantic gap [66], i.e., the disagreement between visual similarities and semantic similarities, causes CBIR to be a very difficult problem.

Early successful attempt by QBIC (Query by Image Content) [48] used very simple visual similarities: quadratic distances between visual features such as color histograms, and thus the discrepancy from semantic similarities was significant. Recent advances in computer vision and multimedia narrowed this gap, as explained in this section, and nowadays researchers have been focusing on couple of specific aspects of semantic similarities in image search by visual example. One is semantic similarity based on the same category of objects or scenes, for example, given a dog image retrieving images of any dogs, and another is based on the same instance of objects or scenes, for example, given a dog image retrieving images of that specific dog. The former is sometimes called concept-based image search, and the latter corresponds to instance search which this paper deals with. Figure 1 shows standard processing flow of instance search with pointers to relevant sections.

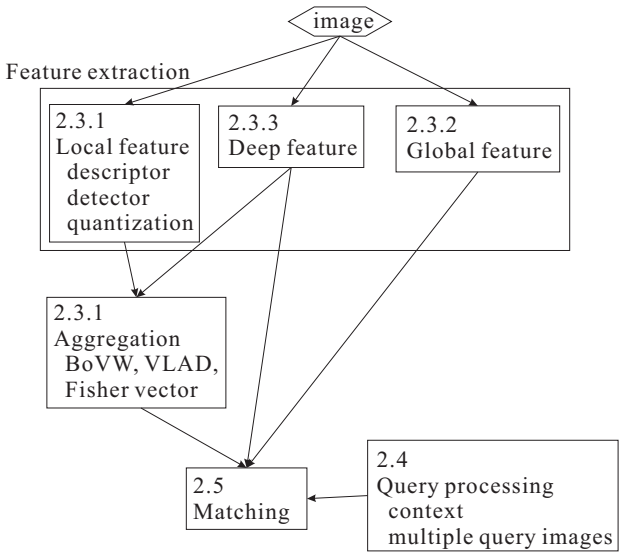


Fig. 1 Standard processing flow of instance search

2.2 Related Benchmarks

Benchmark datasets are used to evaluate the performance of algorithms such as instance search as well as to design the system for fine tuning parameters. Table 1 shows summary of the datasets. COIL-100 [46] is one of the earliest datasets designed for object classification. The dataset is composed of images of 100 specific objects, most of them are commercial products such as candy and medicine packs, from 72 directions (5 degrees apart rotated around a vertical axis) with black background, in total 7200 images. Therefore the dataset, composed of images of instances of 100 different objects and thus one of the earliest object classification datasets, can be regarded as an instance search dataset. Afterwards, to address more challenging situation of object classification, recent datasets for this problem incorporate images of a couple of classes of objects instead of instances of the same objects, such as PASCAL VOC [17] and ImageNet [16].

On the other hand, many datasets especially designed for instance search scenario have been produced and widely used. Typically such datasets contain only specific types of instances such as landmarks and specific objects, partly because technologies required for each type may be different from the others, so that researchers can focus on specific research topics. Most well-known landmark dataset is Oxford Building [54],

Table 1 Instance search datasets

Database Name	Object type	Image composition
COIL-100	specific object	7200 images (100 objects × 72 poses)
Oxford Building 5k	landmark	11 landmarks, 5062 images incl. 5 queries w. ROI per landmark
Paris 6k	landmark	11 landmarks, 6412 images incl. 5 queries w. ROI per landmark
BelugaLogos	logo	26 logos, 10000 images, queries: Qset1 (55 queries, images w. ROI), Qset2 (26 logo thumbnails)
FlickrLogos-32	logo	32 logos, train: 10 hand-picked images per class, val.: 10 images per class + 3000 no logo images, test: the same config. with val.
UKBench	specific object	2550 objects, 4 images per object
SMVS	specific object	8 categories (CD, Book, Landmark, ...), 1200 database images, 3300 query images w. distortion

which is composed of images of 11 different landmarks in Oxford, in total 5062 images including 5 query images per landmark. Paris [55] is another example composed of images of landmarks in Paris. Datasets for logo retrieval are also widely used. BelugaLogos [32, 37] dataset is composed of 10 000 images provided by Belga press agency with global ground truth of 26 different logos (whether a logo present or not) and with local ground truth of 37 logos (with surrounding rectangles). FlickrLogo-32 dataset [60] is composed of images retrieved from Flickr with ground truth of 32 logos. The dataset provides pixel-level annotations which is similar to mask information provided by TRECVID instance search task. As for specific object datasets, UKBench dataset [49] contains 10 200 images of 2550 different objects, 4 images with different conditions for each object. Stanford Mobile Visual Search (SMVS) Data Set [9] contains 1200 images for database, one image per object, and 3300 query images taken by different conditions with mobile devices. The dataset contains images of mostly specific objects but also landmarks (500 landmarks). Face or person datasets, if we regard each individual as a specific object, can be regarded as instance search datasets. Since the history of face recognition research is very long, there are myriad face datasets such as FERRET [56], Multi-PIE [23], Labeled Faces in Wild (LFW) [26], among others.

2.3 Features

2.3.1 Local Features

Local feature descriptors Local features are image features computed in small vicinities of key points, normally aiming at invariance to image scale and rotation, as well as robustness to affine/perspective distortion, viewpoint change, noise, illumination change, background clutter, occlusion, and so on. Scale-invariant feature transform (SIFT) [40] is the best known among such features, and was first designed to match corresponding points for stereo vision. There are many local features proposed following the success of SIFT, however, for instance search scenario, SIFT and its variants such as Speeded Up Robust Features (SURF) [6] and Gradient Location and Orientation Histogram (GLOH) [43] are used in most cases. Although local features are originally not designed to match points between images of the same category, they are intensively used for image categorization and are known to perform well. However, since local features are inherently designed to match corresponding points of the same object observed from different viewpoints, they obviously are more suitable for instance search than for image categorization and image search based on object/scene categories. Original SIFT is designed for monochrome images. However, couple of variations of SIFT which take into account color information are proposed, such as color SIFT and opponent SIFT [61], and are known to be beneficial for instance search scenario especially when color information of objects is distinctive. Since SIFT is essentially histogram (gradient histogram in local region), metrics other than the Euclidean distance, such as χ^2 distance, histogram intersection, and Hellinger distance, may be more appropriate. RootSIFT [3] is known to boost the performance in image retrieval by simply taking the square root of each component of SIFT features to make the Euclidean distances between RootSIFT features compatible with Hellinger distances between SIFT features.

Interest point detector Local features such as SIFT are very discriminative in retrieving the same instances of an object, provided that the local features are computed at exact corresponding points of the object. Therefore, in order to take full advantage of the discriminating power of local features, the proper design of strategies to select feature points (usually called interest points) is very important. There are mainly two strategies to select feature points: sampling-based methods which select feature points without referring to images, and methods using interest point detectors which select feature points referring to images. Sampling-based meth-

ods typically select feature points at every pixel at fixed intervals (e.g., a few pixels) and at multiple scales (multiple sized regions), or select feature points at random locations. If feature points are sampled at extremely high density (e.g., at every pixel), the chance that the corresponding points at the exact corresponding points of an object will become high, however, at the cost of huge number of unmatched points. Feature point detectors, on the other hand, are designed to detect characteristic regions in images, such as corner-like structure, blob-like structure, and so on, hoping that exact corresponding points can be detected even with imaging condition changes such as viewpoint changes.

In an image categorization scenario, it is known that interest point detection is not very helpful, but instead, sampling strategy especially with high density (sometimes called dense sampling) is more effective [50, 75]. On the other hand, in an instance search scenario, interest point detectors are known to be effective (e.g., [54] in matching landmarks).

There are many interest point detectors proposed such as Difference of Gaussian [40], Harris [24, 62], Harris-Affine, Harris-Laplace, Hessian-Affine [42], Maximally Stable Extremal Region (MSER) [41], among others. Extensive comparison can be found in [44] in various aspects such as repeatability. When applied to instance search problem, these detectors have pros and cons, depending on types of objects, and it is also known that the combination of multiple feature point detectors is effective.

Quantization and aggregation As described, local features are very effective for instance search, provided that appropriate matching techniques are used between local features of query images and local features of database images. In searching for matching local features given a query local feature, it is known that the ratio of the distance to the first nearest neighbor and the second nearest neighbor is a very effective criterion [40], however, since typical local features are high-dimensional data (e.g., a SIFT feature is 128 dimensional vector), this requires huge number of nearest neighbor search operations in high-dimensional space and thus this is impractical. For example, assume an image database composed of one million images with one thousand local features for each image. Then given a query image with 1 thousand local features, the image search inherently requires one thousand times nearest neighbor search operations over one billion local features.

In order to significantly accelerate nearest neighbor search, vector quantization using clustering is typically used: local features projected into the same cluster are regarded as matching local features, and otherwise not matching. The number of clusters is a very important design parameter. If very fine clustering is used, matching local features are very close to each other, while the probability that nearby local features fall into different clusters will increase as well (quantization error). On the other hand, if coarse clustering is used, matching local features may not be sufficiently close to each other. For instance search, it is known that very fine quantization (typically one million clusters) is effective [49, 80] despite possible negative impact due to quantization error. Typically used clustering algorithm is k-means [39], however, it tends to be slow both in computation and convergence especially when very large number of clusters are requested. To alleviate this problem, hierarchical k-means (HKM) [49] was used, and now approximate k-means (AKM) [54] is known to perform better because of low quantization error. The implementation of FLANN [45] is also widely used for cluster assignment by fast approximate nearest neighbor search. Hamming embedding [29] is another option: this technique "embeds" binary signature in addition to cluster (voronoi cell) assignment to realize finer quantization. Finer quantization can be achieved by referring to Hamming distances between binary signatures within a cluster.

Quantized local features obtained from each image are then aggregated for image-level representation. Widely used representation is bag of visual words (BoVW) which is employed for image classification [13] and image/video retrieval [65]. This representation regards each cluster as a visual word, and an image composed of multiple local features (thus regarded as multiple visual words) is then represented as a histogram showing occurrences of words. Image similarities are then evaluated by metrics between histograms, e.g., Euclidean distance, Manhattan distance, (other types of) Minkowski distance, χ^2 distance, among others. Typically tf-idf (term frequency-inverse document frequency) weighting or its variants are applied. Since histograms can be regarded as voting by local features, sometimes soft-voting (soft-assignment) is considered, namely, instead of voting only for the corresponding clusters, voting for multiple clusters which are close to the local features. Weights are determined based on distances to cluster centers or rank. Soft-voting is known to be effective in a classification scenario [21] and in instance search as well [55].

Besides BoVW, other aggregation techniques have been proposed such as Sparse Coding [74], Fisher Vector [53], Vector of Locally Aggregated Descriptors (VLAD) [31], etc., and are successfully applied to image classifi-

cation and image retrieval. However, for instance search problem, BoVW approaches are still the most popular.

2.3.2 Global Features

In contrast to local features, global features are features computed for the entire region or significantly large subregion of images. Typical examples include color histogram [67], Color correlogram [27], GIST [52], Local Binary Pattern (LBP) [51], and Histogram of Oriented Gradients (HOG) [14]. Since global features do not require interest point detectors, they may be suitable for images without significant interest points. On the other hand, because of their holistic nature, global features are usually less robust to background change and thus not well suited for instance search. However, if combined with techniques to properly localize the target objects, global features can boost the performance of instance search. For example, Deformable Part Model (DPM) [19] combines HOG with Latent SVM to localize objects.

2.3.3 Deep Features

These days deep convolutional neural networks (DCNN) are successfully applied to many visual tasks including image classification [34]. As this paper suggested, image similarity can be defined by the Euclidean distance between responses of fully connected layers, and many approaches use the responses of fully connected layers of DCNN as semantically rich discriminative features. These features can also be regarded as global features, however, if properly trained, these features can be robust to background clutter (e.g., DCNN trained with large volume of videos can detect cat faces despite the existence of background [35]).

Initial attempts of the application of DCNN to image retrieval were, however, unsuccessful. Babenko et al. [5] report one of the first attempts to use DCNN responses as holistic features of images but the performance is not better than the state of the art local feature-based methods. Researchers then realized that the better performance is achieved when DCNN features are computed at small subregions in images, namely, DCNN features are used as local features, and started investigating how such DCNN features should be aggregated to represent image features. Babenko and Lempitsky [4] reveal that simple sum pooling-based aggregation of DCNN features of patches is superior to other "sophisticated" aggregation techniques such as Fisher Vector and VLAD, which are known to perform better with local features such as SIFT. Razavian et al. [58] compare DCNN responses between patches obtained from queries and patches obtained from database images and computes similarities by taking the maximum over patches of database images then taking the average over patches of queries, without feature aggregation. Tolias, Sicre and Jégou [69] refer to responses of convolution layers (not fully connected layers), max-pooling to obtain region features, and then sum-pooling to obtain image features. There are many other papers on DCNN-based image retrieval that have appeared recently, and including the above mentioned papers, most of them uses relatively "easy" datasets (e.g., Oxford Building and UKBench) for evaluation, and thus their effectiveness on "hard" TRECVID Instance Search dataset is still not extensively explored.

2.4 Query Processing

2.4.1 Context

The given object regions (called as region of interest or ROI) is an important part of the query in instance search. Since regions outside the object regions (called as background regions) do not have visual properties of the object, such background regions are regarded as disturbances.

However, generally objects cannot be totally independent of the scene. For example, cars tend to be observed on roads, birds may appear in the sky, houses may be surrounded by bushes, and so on. Therefore, if properly handled, background information helps in handling object region information. Such background information is called context information.

The usage and the effectiveness of context is well studied for image classification [57,72]. Statistical dependency has been modeled in [57] between object region and context using conditional random field (CRF) and successfully improves the performance of object categorization.

Effective spatial extent of object region to boost object categorization performance has been thoroughly studied in [72]. Its effectiveness for instance search is also confirmed [81] with a couple of instance search datasets including TRECVID Instance Search datasets. Features such as BoVW are obtained both from object regions and background region. And then search results from database using them (note that object regions of images or videos in database are unknown). The final results are then obtained by fusing both lists. Thorough experiments using Oxford, Paris, and TRECVID Instance Search datasets are conducted.

2.4.2 Multiple Examples

If multiple example images are available for a target object, the search performance for the target object can be boosted compared to the case when only one example image is available. Arandjelovic and Zisserman [2] assume such a situation and study a couple of different methods to combine multiple queries using the Oxford building dataset. The paper suggests obtaining independent ranked list for each example image, and combining the ranked lists by using max-pooling (takes maximum score for each image). Zhu, Huang and Satoh [79] also study fusion of multiple example images using Oxford Building as well as TRECVID Instance Search 2011 and 2012 datasets. Interestingly slightly a different conclusion is drawn: this paper advocates average-pooling (averaging scores for each image).

2.5 Matching

2.5.1 Efficiency

The search efficiency is also very important issue especially when the size of the database is huge. As described, the representation based on bag of visual words with very fine quantization is known to be effective for instance search. In this situation, each image is represented as a very sparse histogram of visual words, and this is very similar to bag of words representation for text. Therefore, efficient indexing techniques developed for text retrieval are applied for visual object retrieval and instance search.

The most well known and frequently used technique is the inverted index (inverted file) [84] where a look-up table is prepared for each (visual) word to quickly find documents (images/videos) containing the (visual) word. The inverted index was applied to visual search in a very early attempt [65]. Min-hash [8] is based on multiple hash functions corresponding to multiple permutated and numbered vocabularies. Each hash function returns the minimum value for each permuted vocabulary. The search is accelerated based on the fact that the probability that hash values of two documents agree converges to the similarity of the documents (in Jaccard similarity). A couple of attempts can be found to apply min-hash to visual search [12, 10, 59].

On the other hand, some representations of images or videos other than bag of visual words may not be sparse but dense vectors such as Fisher Vector and VLAD. Product quantization (PQ) [30], which is based on quantizing subvectors, is known to perform well in accelerating search based on dense vector representations.

2.5.2 Geometric Consistency

In instance search, relevant images in database should contain the same instances of object as the query. Therefore, if we compare query images and relevant images, they should share the same instances of the target objects.

In this case, they are likely to share corresponding surfaces of the object instances, and thus there will be likely a dense patch-wise (or point to point) correspondences between query images and relevant images. Obviously corresponding point pairs yield some kinds of geometric consistency, and thus it is known that geometric consistency checking may boost the performance of instance search. An example of geometric consistency is homography: when query and relevant images share planar corresponding surface (in reality any surface can be regarded as piecewise planar), corresponding point pairs are related by a homography. Given point correspondences (normally obtained by interest point detector and point matching by local features), Random sample consensus (RANSAC) [20] effectively finds homography on which largest number of point correspondences agree by random sampling and iterative consistency checking. RANSAC is originally developed for binocular stereo vision, but effectively applied to instance search problem as a post processing [11]. Variants of RANSAC for instance search are also proposed (e.g., LO-RANSAC [36]).

RANSAC is known to be slow due to random sampling and iteration. To speed up the geometric consistency checking for instance search by using the idea similar to Hough transform, weak geometric consistency (WGC) checking [29] is proposed. WGC can effectively filter out irrelevant local descriptors, and can be integrated into an inverted file for efficient retrieval. Other techniques which embed geometric information into local features and integrated into indexing mechanism taking into account both patch-wise local appearances and geometric information have been proposed. [76] applies Delaunay triangulation to interest points in each image to generate a planar graph, and retrieve images corresponding to graphs having similar structure to a query. Geometric min-Hash [10] uses central features and indexes them using min-Hash, similar to the standard BoVW framework, and also uses secondary features for each central feature which can be found in neighborhood of the central feature with similar scale. By using the similarity in local feature space for both central feature and secondary feature, Geometric min-Hash guarantees geometric constraint among pairs of interest points and boosts the performance of instance search.

Bundled features [73] uses a similar idea: bundle multiple interest points in a local neighborhood, use them together to describe the region, and incorporate them into an inverted file. Geometry-preserving visual phrases [77] encodes not only local vicinity but also long-range spatial layouts by using offset space describing relative spatial locations of pairs of interest points. The information is shown to be integrated into min-Hash. [59] encodes spatial layout into two indices: the primary index describes pairs of local features, and for entries found in the primary index, the secondary index will be searched which describes triples of local features. [38] embeds spatial layout into an inverted file by using spatial context based on spatial relationship dictionary which encodes patch-wise appearance and relative location of pairs of local features, followed by binary signature encoding the spatial context.

3 TRECVID Data

There have been two primary, related difficulties in evaluating instance search systems in TRECVID: finding realistic data with sufficient repeated instances and then creating realistic test topics that fit the data. For three years TRECVID experimented with three very different sorts of data before beginning in 2013 with a larger, better-suited dataset that could support at least three additional years of evaluations. Figure 2 shows sample frames from the different datasets used between 2010 to 2015. The primary decision making factor in selecting those datasets where the ability to find recurring materials of specific instances for people, objects, and locations.

2010 - Sound and Vision: In 2010 professionally created video from the Netherlands Institute for Sound and Vision was used (≈180 h in MPEG-1 format). Recurring news programming offered repeated instances of politicians and locales. Several sketch comedy programs for children contained actors that appeared over and over as the same characters but in different clothing and settings. Sports reporting included logos. The video was automatically divided into 60 000 shots.

NIST (National Institute of Standards and Technology) staff watched a subset of the test videos and created eight topics looking for video of a character, another eight looking for an individual person, an equal number targeting objects, and one asking for video of a location - for a total of 22 topics. Each topic contained about five example images, each with a rough binary polygonal mask marking the position of the topic target in the image and a set of (x,y) coordinates for the vertices of the mask.

2011 - BBC travel rushes: In 2011 unedited video intended for BBC travel programming was used (≈81 h in MPEG-1 format). Presenters recurred, as did varying views of particular buildings, animals, architectural details, vehicles, etc. The videos were divided automatically into 10 491 shots of fixed length. Since the number of test shots was relatively small, an attempt was made to supplement these by adding variants to simulate video of the same target but from a different angle, using a different camera, in different lighting. To this end a copy of each original test shot was transformed in a randomized fashion with respect to gamma, contrast, aspect, and hue and then added to the test set to yield 20 982 test shots.

NIST staff watched a sample of the test videos and created 25 topics that targeted objects (17), persons (6) or locations (2). Each topic contained about five examples images with associated masks; the coordinates of the vertices were dropped as participants found them redundant.

2012 - Flickr Creative Commons: In 2012 the evaluation turned for test data to Internet video available for research under a Creative Commons license from Flickr (≈200 h in webm format). Robin Aly at the University of Twente created five sorts of Flickr queries using externally sourced lists of possible targets in the following categories: buildings, castles, events, heritage, and person. These were designed to return repeated shots of the same object, person, or location from multiple sources, e.g., one looking for shots of the Eiffel Tower, the Puma logo, Stonehenge, etc. The videos were automatically divided into 74 958 shots of fixed length.

The search results were reviewed by NIST, 21 search targets were selected, and corresponding topics were created by NIST staff - 15 against objects, 5 locations, and 1 person. Each topic contained about five examples images with associated masks.

2013, 2014, 2015 - BBC EastEnders soap opera: Impressed with the difficulty of finding appropriate instance structure in various videos of the real world, the organizers began early in 2012 to work with the BBC (Andy O'Dwyer) and the Access to Audiovisual Archives (AXES) project (Robin Aly at the University of Twente and Noel O'Connor at Dublin City University) to make video from the BBC soap opera series, EastEnders, available as test data in 2013 (≈464 h in mp4 format). The idea, suggested already in 2010 by Werner Bailer from Joanneum Research (JRS), was to exploit the structure of the small world created for a television series with its recurring, varying objects, people, and interior/exterior locations. The BBC kindly

Fig. 2 Example frames from the different datasets used

provided 244 weekly omnibus videos from 2007 to 2012. These videos present a slowly changing set of recurring people (several dozen), locales (homes, workplaces, pubs, cafes, restaurants, open-air market, clubs, etc.), objects (clothes, cars, household goods, personal possessions, pets, etc.), and views (various camera positions, times of year, times of day). The videos were automatically divided into 471 523 shots.

4 TRECVID Query topic development

NIST staff viewed more than 10 % of the videos chosen at random and made notes about recurring objects, people, and locations. Approximately 90 potential search targets were chosen. Half the object targets were stationary - here the background could be a decisive clue; not so for the mobile objects whose background changed. Topic targets were selected to exhibit several kinds of variability - inherent (boundedness, size, rigidity, planarity), locale (multiplicity, variability, complexity), and camera view (distance, angle, lighting).

For 2013, NIST created 30 topics, a representative one-third sample of the 90 with 26 looking for objects and 4 for people. Half of the person topics were looking for named characters, half for unnamed extras. Each topic contained 4 image examples taken from the test collection. Shots containing the example images were ignored in the scoring. Associated with each example image was a binary mask indicating with a rough polygonal mask where the topic target was located in the image. Also provided was the video shot from which each example image was taken. Participants could indicate with each submission which subset of example images was used and/or whether the video examples were exploited.

Consideration of issues concerning the definition of the masks gradually converged on a set of rules motivated by ease of use for the assumed searcher. For each frame image the binary mask of the region of interest (ROI) was bounded by a single polygon. Where multiple targets appeared in the image only the most prominent was included in the ROI. The ROI could contain non-target pixels, e.g., non-target regions visible through the target or occluding regions.

Topic targets were selected to exhibit several kinds of variability expected a priori to interact with the search engine algorithms and affect the overall effectiveness of the search systems. Search targets with fixed locations may be detectable using the background, while mobile targets will not. Even the same static target will appear differently from shot to shot as the camera/lense position vary and the mobile constituents of the scene change. Variability in the appearance of rigid targets should be less than that of flexible ones. All other things being equal, relatively small targets should provide less information to the detection process than larger ones. Planar targets will likely offer fewer different views than 3-dimensional ones.

In addition to the stationary [S] versus mobile [M] distinction, four simple mutually exclusive topic categories based on the foregoing thinking were used to gauge the diversity of the topics during the selection/creation process:

- A rigid non-planar small
- B rigid non-planar large (> 2 ft. tall)
- C rigid planar, logo
- D non-rigid non-planar (e.g., person, animal, garment, paper)

Table 2 Counts of instance search topics by categories

Year	M	S	A	B	C	D
2013	16	14	8	10	6	6
2014	18	12	7	8	7	8
2015	15	15	11	7	6	6

Table 2 depicts the distribution of topic types for the EastEnders data. See Tables 3, 4, and 5 for a complete listing of EastEnder topics and their types, including the topic number, the year used, the type (*O*bject, *P*erson, *L*ocation), whether *S*tationary or *M*obile, and the category (A,B,C,D) as listed above.

We can formulate four simple-minded expectations in terms of the above categories and based on the notion that greater variability of targets generally results in harder topics. If we rank topics by the mean effectiveness across all systems then we would expect to find category S topics generally ranked higher than category

M topics. This and other expectations can be formulated as follows:

1. S > M (stationary should be easier than mobile)
2. B > A (larger should be easier than smaller)
3. C > A, B (planar should be easier than non-planar)
4. A, B, C > D (rigid should be easier than flexible)

Table 3 EastEnders topic information in 2013

Topic	Type	S/M	Cat	Text
9069	O	S	C	a circular 'no smoking' logo
9070	O	S	B	a small red obelisk
9071	O	M	C	an Audi logo
9072	O	M	C	a Metropolitan Police logo
9073	O	S	A	this ceramic cat face
9074	O	M	A	a cigarette
9075	O	M	A	a SKOE can
9076	O	S	B	this monochrome bust of Queen Victoria
9077	O	M	D	this dog
9078	O	S	C	a JENKINS logo
9079	O	S	B	this CD stand in the market
9080	O	S	B	this public phone booth
9081	O	M	B	a black taxi
9082	O	M	C	a BMW logo
9083	O	M	A	a chrome and glass cafetiere
9084	P	M	D	this man
9085	O	S	C	this David refrigerator magnet
9086	O	S	B	these scales
9087	O	M	C	a VW logo
9088	P	M	D	Tamwar
9089	O	M	A	this pendant
9090	O	S	B	this wooden bench with rounded arms
9091	O	M	A	a Kathy's menu with stripes
9092	P	M	D	this man
9093	O	S	B	these turnstiles
9094	O	M	A	a tomato-shaped ketchup dispenser
9095	O	S	B	a green public trash can
9096	P	M	D	Aunt Sal
9097	O	S	A	these checkerboard spheres
9098	O	S	B	a P (parking automat) sign

Table 4 EastEnders topic information in 2014

Topic	Type	S/M	Cat	Text
9099	O	M	C	a checkerboard band on a police cap
9100	O	M	A	a SLUPSK vodka bottle
9101	O	S	B	a Primus washing machine
9102	O	S	B	this large vase with artificial flowers
9103	O	M	A	a red, curved, plastic ketchup container
9104	P	M	D	this woman
9105	O	M	D	this dog, Wellard
9106	O	S	C	a London Underground logo
9107	L	S	B	this Walford East Station entrance
9108	O	S	A	these 2 ceramic heads
9109	O	M	C	a Mercedes star logo
9110	O	S	B	these etched glass doors
9111	O	S	C	this dartboard
9112	O	S	C	this HOLMES lager logo on a pump handle
9113	O	M	D	a yellow-green sanitation worker vest
9114	O	S	B	a red public mailbox
9115	P	M	D	this man
9116	P	M	D	this man
9117	O	S	A	this pay phone
9118	O	M	C	a Ford Mustang grill logo
9119	P	M	D	this man
9120	O	S	B	a wooden park bench, straight-backed, with flat arm rests
9121	O	M	D	a Royal Mail red vest
9122	O	M	A	this round watch with black face and black leather band
9123	O	M	A	a white plastic kettle with vertical blue window
9124	P	M	D	this woman
9125	O	M	B	this wheelchair with armrests
9126	O	M	C	a Peugeot logo
9127	O	S	B	this multicolored bust of Queen Victoria
9128	O	M	A	this F pendant

The topic process from 2013 was continued without major change in 2014 and 2015 using new subsets of the 90 potential search targets. This allowed us to measure participating systems performance without introducing significant changes each year.

Table 6 presents the basic information on the data used in TRECVID: the year, the data source (Sound & Vision, BBC rushes, Flickr Creative Commons, BBC EastEnders), the number of test shots, the average shot duration in seconds, the number of test topics, how many topics targeted objects, persons, and locations, as well as what percent of the test shots were found responsive to a topic (true positives). As can be seen, the task has focused increasingly on objects. Participants early on expressed a desire not to emphasize search for persons as it was felt this might be dominated by face matching which receives attention in other venues. Searching for locations presents special problems because the target of the search is so large that the variety of views is enormous. In addition, very large objects can be seen as locations if a person can move into, around, under, or above them, e.g., the Eiffel Tower, Stonehenge, Prague Castle.

Table 5 EastEnders topic information in 2015

Topic	Type	S/M	Cat	Text
9129	O	M	A	this silver necklace
9130	O	M	A	a chrome napkin holder
9131	O	M	A	a green and white iron
9132	O	M	A	this brass piano lamp with green shade
9133	O	M	A	this lava lamp
9134	O	M	A	this cylindrical spice rack
9135	O	M	B	this turquoise stroller
9136	O	M	B	this yellow VW beetle with roofrack
9137	O	M	C	a Ford script logo
9138	O	M	D	this man with moustache
9139	O	M	D	this shaggy dog (Genghis)
9140	O	M	D	a Walford Gazette banner
9141	O	M	D	this guinea pig
9142	O	M	D	this chihuahua (Prince)
9143	P	M	D	this bald man
9144	O	S	A	this doorknocker on #27
9145	O	S	A	this jukebox wall unit
9146	O	S	A	this change machine
9147	O	S	A	this table lamp with crooked body
9148	O	S	A	this cash register (at the cafe)
9149	L	S	B	this Walford Community Center entrance from street
9150	O	S	B	this IMPULSE game
9151	L	S	B	this Walford Police Station entrance from street
9152	O	S	B	this PIZZA game
9153	O	S	B	this starburst wall clock
9154	O	S	C	this neon 'Kathys' sign
9155	O	S	C	this dart board
9156	O	S	C	a 'DEVLIN' lager logo
9157	O	S	C	this picture of flowers
9158	O	S	C	this flat wire 'vase with flowers'

Table 6 Overview of TRECVID Instance Search Data

Year	2010	2011	2012	2013	2014	2015
Source	SV	Rush	Flickr	EE	<–	<–
Shots	60k	21k	75k	471k	<–	<–
Avg duration	11	28	10	3.5	<–	<–
Topics	22	25	21	30	27	30
Object	8	17	15	26	21	27
Person	13	6	1	4	5	1
Location	1	2	5	0	1	2
TPs	2.0	8.7	1.6	2.9	2.8	2.6

5 Overview of INS task results (2010-2016)

This section summarizes the results of systems evaluation the last six years in two parts. Between 2010 to 2012, three unique datasets were used in pilot evaluations so that comparison of systems across years would be confounded with the effect of changing data. However, in 2013 to 2015 the BBC Eastenders system scores

Fig. 3 Examples of pilot evaluation topics (Objects, Persons, Locations)

Table 7 Overview of TRECVID Automatic Instance Search Results - Maxima and Means

Year	Obj_max	Per_max	Loc_max
2010	0.095	0.5	0.129
2011	0.960	0.723	0.921
2012	0.717	0.761	0.820
2013	0.860	0.439	-
2014	0.977	0.167	0.382
2015	0.911	0.701	0.856

Year	Obj_mean	Per_mean	Loc_mean
2010	0.003	0.012	0.014
2011	0.160	0.134	0.292
2012	0.063	0.407	0.132
2013	0.106	0.058	-
2014	0.179	0.023	0.118
2015	0.205	0.074	0.283

are comparable since the same testing data was used with different but very similar sorts of topics for each year. Examples of the selected topics in pilot years can be shown in Figure 3 while some of the topics used between 2013 to 2015 can be shown in Figures 4 to 6. We summarize here the effectiveness scores per topic and per topic category (Objects, Persons, Locations) for automatic and interactive runs, the relation between the system scores and processing time, and the relation between per-topic scores and number of found true positives. (In 2010 an extra topic type, "Characters", was distinguished from "Persons", but not in subsequent years).

A summary of the best and mean scores for each topic type across all years is shown in Tables 7 and 8

Fig. 4 2013: Examples of BBC Eastenders topics (Objects, Persons)

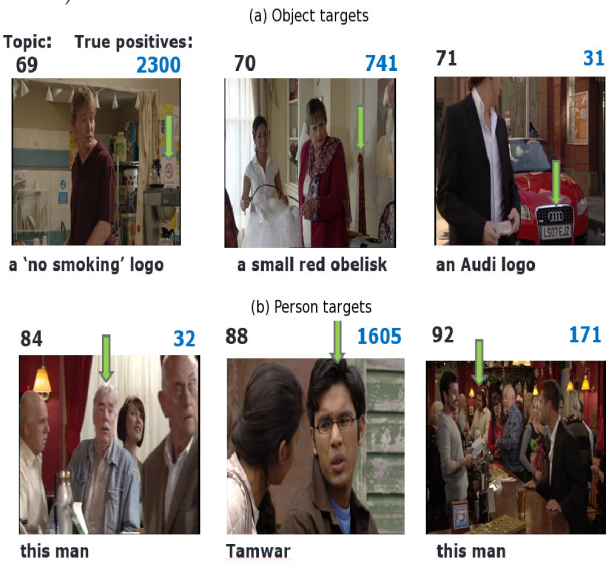
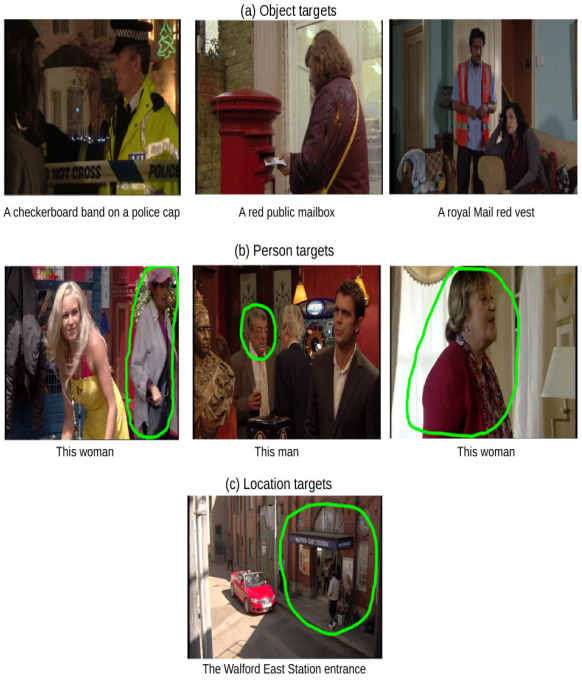


Fig. 5 2014: Examples of BBC Eastenders topics (Objects, Persons, Location)



for automatic and interactive runs respectively. The relation between MAP and processing time between 2011 to 2015 is shown in Figures 7 to 11, while Figures 12 to 16 show the relation between maximum AP and number of found true positives. More detailed results for each of the pilot and BBC Eastenders evaluation years are summarized in the next sections followed by observations.

Fig. 6 2015: Examples of BBC Eastenders topics (Objects, Persons, Location)

Table 8 Overview of TRECVID Interactive Instance Search Results - Maxima and Means

Year	Obj_max	Per_max	Loc_max
2010	-	-	-
2011	0.743	0.564	0.771
2012	0.633	0.711	0.831
2013	0.663	0.550	-
2014	0.944	0.176	0.305
2015	0.892	0.183	-

Year	Obj_mean	Per_mean	Loc_mean
2010	-	-	-
2011	0.257	0.172	0.569
2012	0.119	0.509	0.205
2013	0.121	0.10	-
2014	0.158	0.031	0.093
2015	0.220	0.094	-

5.1 2010-2012: Three pilot evaluations

The three years pilot evaluations helped the organizers to refine the instance search task with its possible topic types and helped the participant systems to better get sense of what to expect when asked to search for specific video instance using few examples and almost unconstrained testing video collections.

In the first year, 15 research teams submitted 39 runs. 8 object topics, 13 person topics (including characters) and 1 location topic were created by NIST from the sound & vision video data. The top half of the runs

Fig. 7 2011: MAP vs. elapsed time

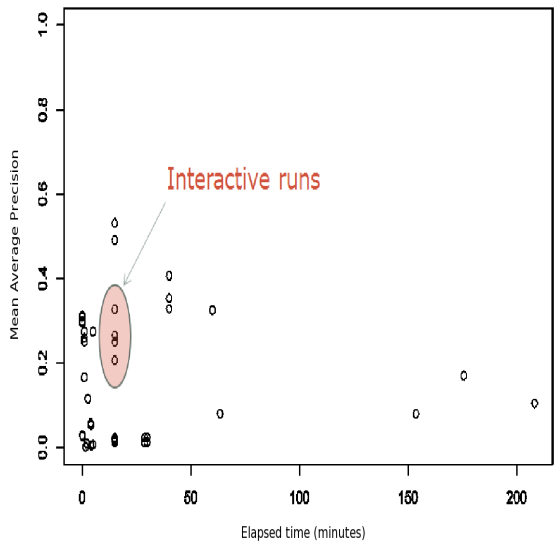


Fig. 8 2012: MAP vs. elapsed time

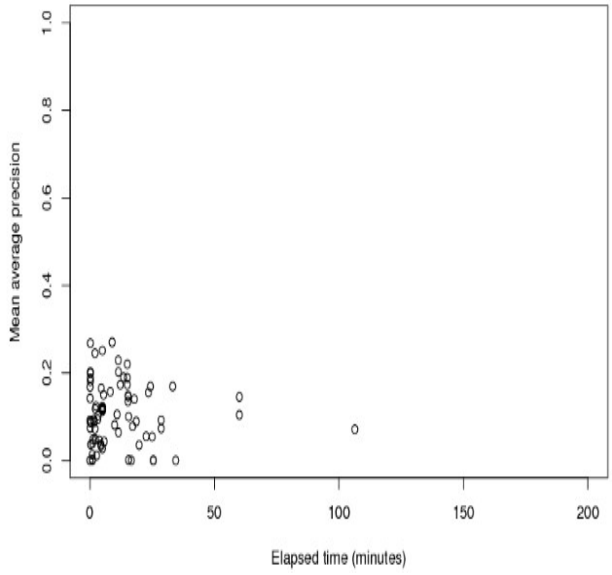


Fig. 9 2013: MAP vs. elapsed time

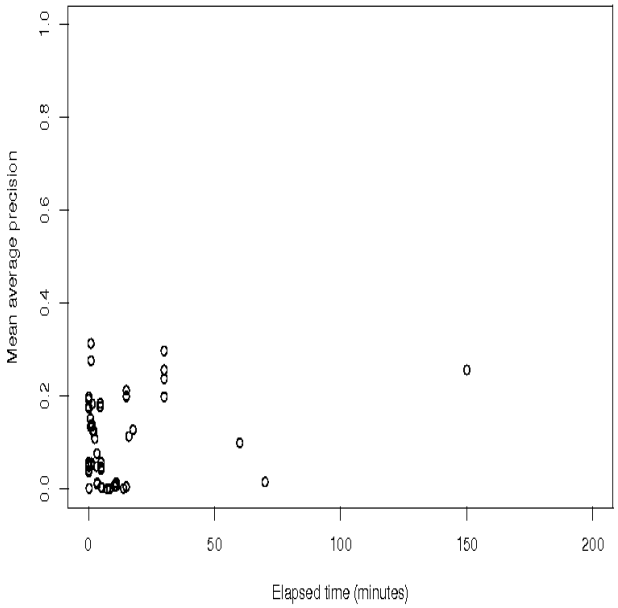
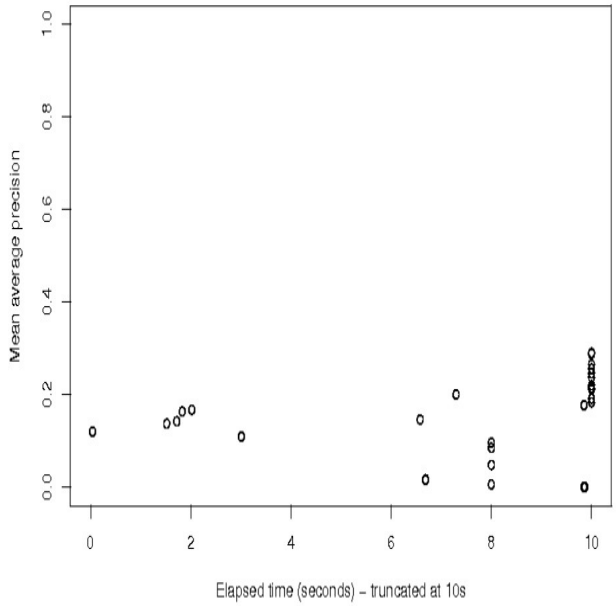


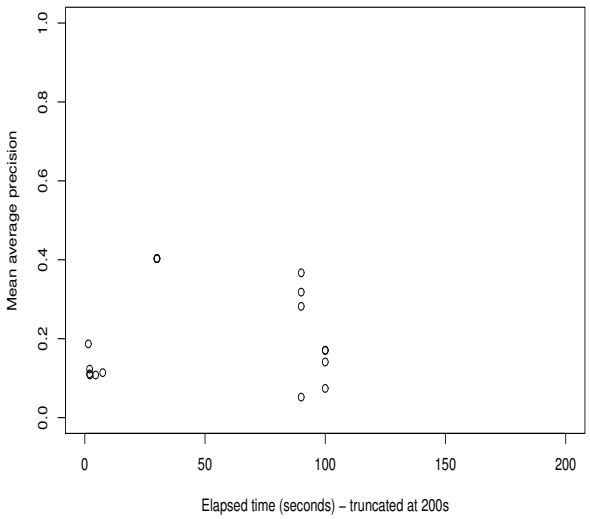
Fig. 10 2014: MAP vs. elapsed time for fastest runs (in seconds)



had mean average precision (MAP) scores ranging from 0.01 to 0.03. Figure 17 depicts the distribution of scores by topic for the people, character, location, and object types. During the TRECVID 2010 Workshop there was a panel discussion out of which came the suggestion that if we continued to use small targets, then we should use better quality video. In general, results of this year were of a very preliminary nature with very low MAP scores and a lot of topic type specific approaches.

In the second year of the pilot task, 13 research groups submitted 37 automatic runs and 4 interactive runs. Overall, 17 object topics, 6 person topics and 2 location topics were created from the BBC rushes dataset by NIST. Figure 18 is a boxplot showing the distribution of effectiveness scores (average precision) by topic and topic type, as achieved by fully automatic systems. Figure 19 provides the corresponding information for interactive runs. Figure 20 shows scores of the top-scoring runs. Surprisingly, some fully automatic

Fig. 11 2015: MAP vs. elapsed time



runs achieved better effectiveness than most of the interactive runs. An analysis to the submitted results per topics show that teams treated the original and transformed clips independently. Although, in general absolute results per topic type are better than previous year, we can not directly compare the two years because the two datasets are different.

In the third pilot year of the task, 24 teams submitted 79 automatic and 15 interactive runs. Figures 21 and 22 are boxplots showing the distribution of per-topic average precision scores across all automatic and interactive runs for each topic type respectively. The test collection size is too small to draw strong conclusions about the differences due to topic type. Comparing the best performance by topic in interactive versus automatic runs, Figure 23 shows progress for interactive runs where they outperformed automatic ones on 8 of the 21 topics compared to 2011 (2 of the 25 topics).

To summarize our observations for pilot years, first, systems scored best on locations, where they can use the entire frame. Specifically, in 2012 the set of location topics targeted popular locations (e.g. Prague Castle, Hagia Sophia, Hoover Dam, Pantheon, Stonehenge) with very unique appearance. Second, more processing time was not necessarily required for better scores and many fast systems achieve same or better performance compared to slower systems (as shown in Figures 7 and 8). Third, no clear correlation was found as one might expect between AP and number of found true positives (Figures 12 and 13) which may indicate that the systems did not invest too much in developing sophisticated (re)ranking strategies to boost their performance. Finally, although it is hard to compare systems across those three years, perhaps the most common observation is that there was big variation across topic performance in general and within each topic type.

Fig. 12 2011: True positives per topic vs. maximum average precision

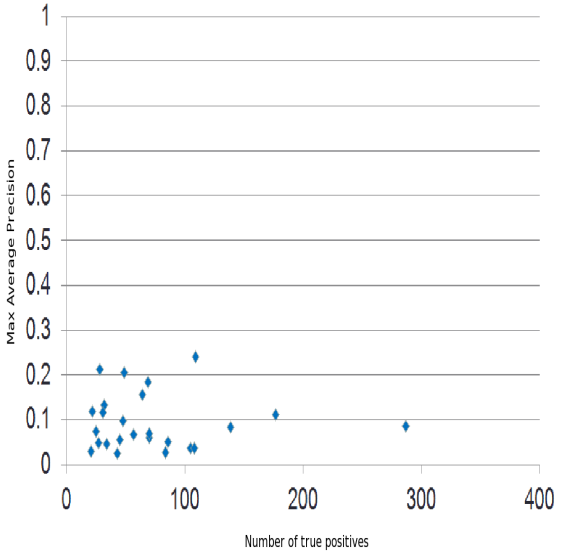


Fig. 13 2012: True positives per topic vs. maximum average precision

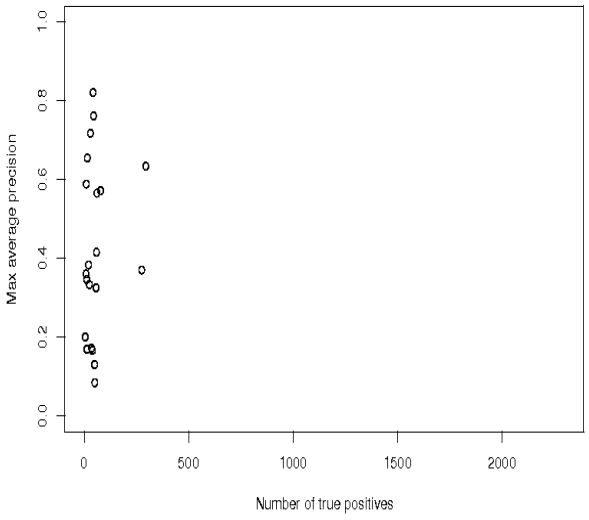


Fig. 14 2013: True positives per topic vs. maximum average precision

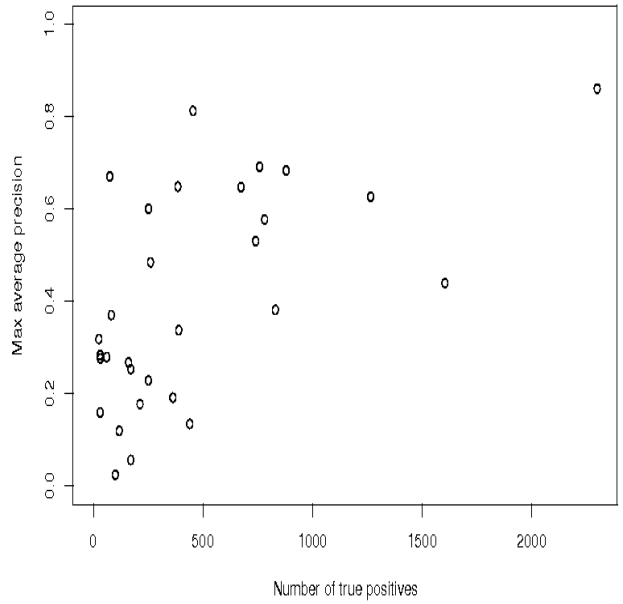


Fig. 15 2014: True positives per topic vs. maximum average precision

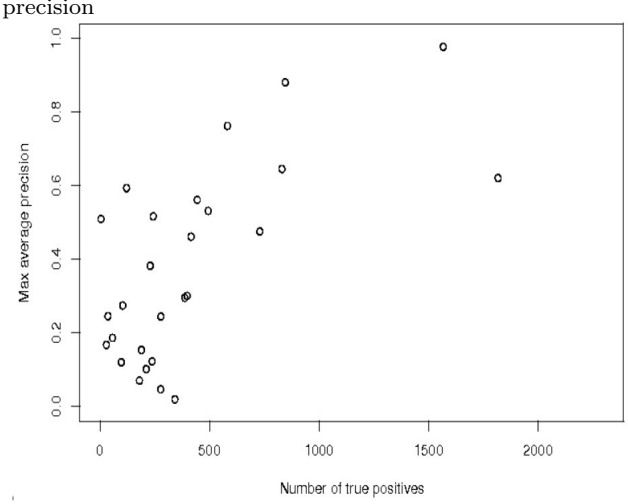
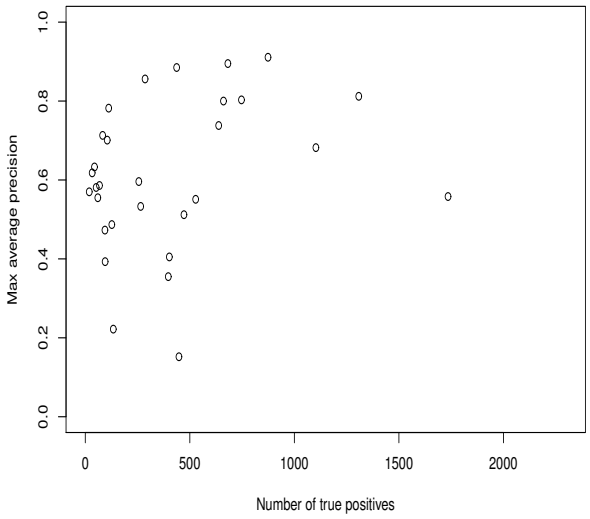


Fig. 16 2015: True positives per topic vs. maximum average precision



5.2 2013 - 2015: The small world of the BBC EastEnders series

During the years of 2013 to 2015, the availability of the BBC Eastenders video dataset allowed the organizers to formulate better the design of the topic categories and exploit the range of available instances within the videos from large locations to small objects giving the opportunity to measure the effect of topic characteristic on the overall system performance. Since the collection to be searched was the same in all three years and the topics were balanced samples from a single larger set, comparison of systems across years is possible.

5.2.1 2013 evaluation

In 2013, using the BBC Eastenders videos dataset, 22 groups submitted 65 automatic runs and 9 interactive runs (using only the first 24 topics). 26 object topics and 4 person topics were selected by NIST for this year.

Figure 24 shows the distribution of automatic run scores (average precision) by topic as a boxplot. Topics are sorted by maximum score with the best performing topic at the left. Median scores vary from about 0.3 down to almost 0.0. Per topic variance varies as well with the largest values being associated with the topics that have the best performance.

In Figure 25, a boxplot of the interactive runs' performance, the best median is actually slightly below that for the automatic runs. Topics with targets that are stationary, rigid objects make up 5 of the 12 with the best scores, but such targets also make up 4 of the bottom 12 topics.

Easier topics seemed to be the ones with simple visual context, stationary targets, or planar and rigid objects. While more difficult topics tend to be associated with small, or moving targets with different camera angle and locations, or non-planar and non-rigid objects.

5.2.2 2014 evaluation

In the second year of using the Eastenders dataset 23 groups submitted 107 automatic and 12 interactive runs (using only the first 24 topics). In total, 27 topics were evaluated including 21 objects, 5 persons and 1 location.

Figure 26 shows the distribution of automatic run scores (average precision) by topic as a boxplot. The topics are sorted by the maximum score with the best performing topic on the left. Median scores vary from nearly 0.8 (higher than 2013) down to almost 0.0. Per-topic variance varies as well with the largest values being associated with topics that had the best performance. The persons topics were the most difficult probably due to the high variability of the appearance of the persons and/or their context.

In Figure 27, a boxplot of the interactive runs performance, the relative difficulty of several topics varies from that in the automatic runs but in the majority of cases is the same. Here, unlike the case with the automatic runs, stationary, rigid targets are equally represented (5 of 11) in the top and bottom halves of the topic ranking.

For topics with less than 500 true positives there seems to be little correlation with effectiveness (See Figure 15). While for those with more than 500 true positives, maximum effectiveness seems to rise with the number of true positives. However, this observation is not obvious in the 2013 results (Figure 14).

Figure 28 shows the relationship between the number of topic example images used and the effectiveness of the runs. (Scores for multiple runs from a team with the same number of image examples used were averaged.) With few exceptions, using more image examples resulted in better effectiveness. However, using the video associated with each image example did not produce any improvement in effectiveness over using just all four image examples. This was the first year video for the images examples was made available and we expected more experiments need to be done by systems to exploit the video example.

5.2.3 2015 evaluation

In the third year 14 groups submitted 44 automatic and 7 interactive runs. Each interactive search was limited to 15 minutes. NIST evaluated 30 topics (from which, 24 topics were for interactive runs) including 26 objects, 2 persons and 2 locations.

Figures 29 and 30 show the distribution of automatic and interactive run scores (average precision) by topic as a boxplot respectively. The topics are sorted by the maximum score with the best performing topic on the left.

Median scores vary from nearly 0.5 down to 0.0 for automatic runs. while interactive runs median scores range from 0.44 down to 0.0. Per-topic variance varies as well with the largest values being associated with topics that had the best performance. For the majority of topics, the relative difficulty seems to be similar between automatic vs interactive runs.

Analyzing the relation between the results and topic difficulties it can be shown that for automatic runs 10 out of the 15 top ranked topics were stationary while 5 out of 15 bottom ranked topics were stationary. The opposite is true for mobile targets. Only 5 out of 15 were among the top 15 ranked topics while 10 were among the bottom 15 ranked topics. Small and rigid targets were harder as only 3 out of 15 were among the top 15 topics while 8 out of 15 were among the bottom 15 topics.

Similarly for interactive results, 7 stationary and 5 mobile targets out of 12 were among the top ranked topics. While 2 stationary and 10 mobile targets out of 12 were among the bottom ranked topics. Unlike the case with the automatic runs, rigid small targets are approximately equally represented (5 of 12) in the top and (4 of 12) in the bottom halves of the topic ranking. Non-rigid non-planar targets were harder as 1 of 12 were among the top halve ranked topics vs 5 of 12 in the bottom halve of ranked topics.

The relationship between the two main measures - effectiveness (mean average precision) and elapsed processing time is depicted in Figure 11 for the automatic runs with elapsed times truncated to 200 s. It can be shown that runs that took long processing times were not necessary better than fast ones and the best performance took 30 s per topic.

The relationship between the number of true positive and the maximum effectiveness on a topic is shown in Figure 16. Similarly to 2014 results, for topics with less than 500 true positives there seems to be little correlation; for those with more than 500 true positives, maximum effectiveness seems to rise with the number of true positives except for couple of topics. In fact analyzing those 9 topics with more than 500 true positives, we found that 8 out of 9 are considered stationary topics in 2015. However the same is not true in 2014 as only 4 out of 9 topics where stationary and has more than 500 true positives. Perhaps systems enhanced their ranking strategies in 2015.

Figure 31 shows the results of automatic runs and distinguishing the ones that used images only examples vs the ones that used video examples plus optionally image examples. Although the top two runs exploited

the video examples, still most submitted runs are just using the image only examples. Clearly the usage of video examples still needs more research from participants.

5.2.4 2013-2015 results summary

In general, within the Eastenders dataset, object topics scores are higher than person topics. This may be due to the fact that chosen objects are unique instances within the videos and in some cases have strong correlation with certain context, background, or characters. On the other hand, although chosen people instances are by definition unique, there is a lot of complexity that systems can face analyzing all characters in the foreground and background of the videos. In addition, different factors can be expected to affect system performance. For example, some topic types may be more difficult than others due to their characteristics (size of region of interest, variability, background complexity, stationary vs mobility, rigidity, planarity), capturing factors (camera angle, lighting, & zoom), frequency of true positives either in training examples or testing dataset, and advancements of used approaches per topic types (face recognition vs object detection).

In regard to the relation between processing time and AP, our observation is similar to pilot years in which more processing time is not necessarily required for better scores and many fast systems achieve same or better performance compared to slower systems (as shown in Figures 9, 10, and 11).

Also, similar to pilot years, no clear correlation is apparent as one might expect between AP and number of found true positives (Figures 12 ,13 and 16) except in some cases when true positives exceed certain threshold. This may indicate that systems still need to develop better ranking strategies to boost their performance.

5.2.5 Influence of topics on performance

In order to test our earlier hypotheses (p. 10) about which topic categories should be easier than others, it was necessary to define a measure for topic easiness to fit our purpose. First we sorted topics by their median effectiveness across all systems.

Then for each hypothesis that a topic category X is easier than category Y we calculate the average value of the ratio between number of times each topic category X is ranked above any topic of category Y to the total number of category Y topics:

$$Easiness(X,Y) = average(\frac{\#N_x}{\#Y_n}) \quad (1)$$

Table 9 Topic categories hypothesis testing results

Hypothesis	2013	2014	2015
S > M	0.75	0.86	0.88
B > A	0.55	0.63	0.66
C > A	0.78	0.71	0.71
C > B	0.70	0.55	0.50
A > D	0.58	0.88	0.75
B > D	0.63	0.94	0.83
C > D	0.77	0.73	0.94

where N_x is number of times a topic in category X is ranked above any topic of category Y, and Y_n is the number of topics in category Y.

The higher the easieness value, the easier X is than Y. In general we consider a topic category X to be easier than topic category Y if the average easiness value is greater than 0.5. Table 9 shows the results of this experiment. Conclusions from results are consistent across years and support the hypotheses that stationary, larger, planar and rigid targets are easier to find than mobile, smaller, non-planar, and flexible ones - although less strongly for some cases (2013:B>A, 2013:A>D, 2014:C>B, 2015:C>B) than others. An additional summary presentation of the raw data distributions for each type confirms that stationary topics are easier than mobile (Figure 32). This distinction explains most of the variability between topic scores. When looking at the different types of topics (small, large, planar, non-planar, non-rigid) the data reveals some patterns: for stationary topics, type C (planar, logo) seems easier than the non-planar types A and B. For the mobile topics, there seems to be no consistent rank order in difficulty between topic types (Figures 33, 34, 35).

Figure 36 shows a sample query from each year from those who achieved the lowest median AP across all runs. From the samples it can be shown that persons, small objects and animals were hard to detect.

Fig. 17 2010: Average precision for automatic runs by topic/type

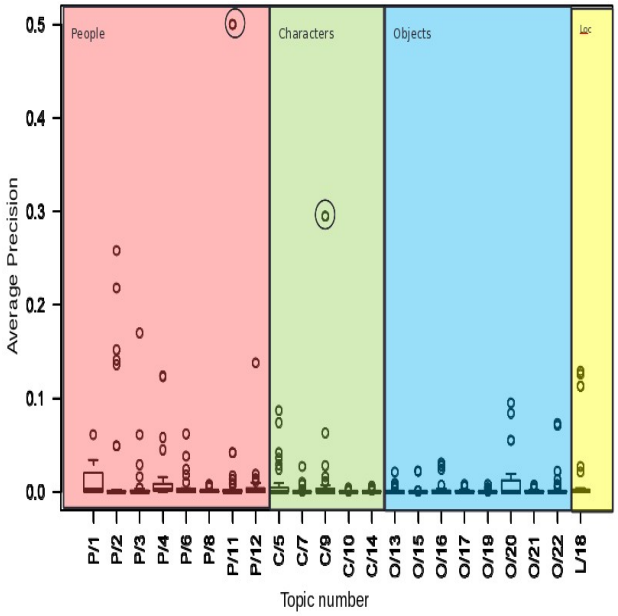


Fig. 18 2011: Average precision for automatic runs by topic/type

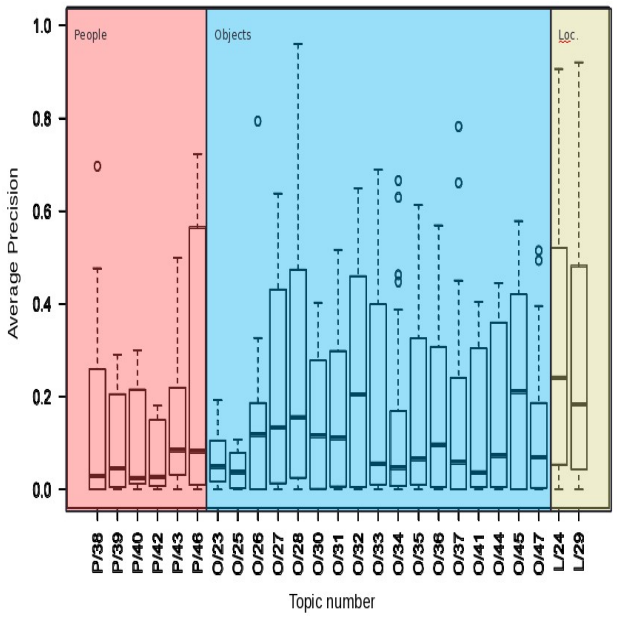


Fig. 19 2011: Average precision for interactive runs by topic/type

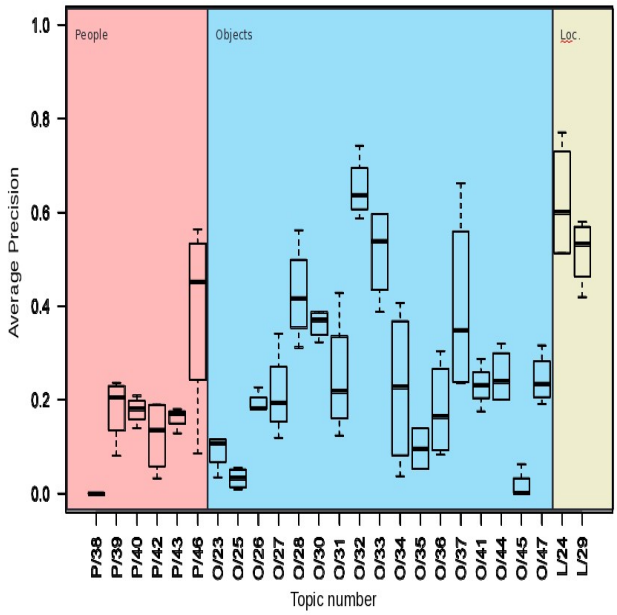


Fig. 20 2011: AP by topic for top runs

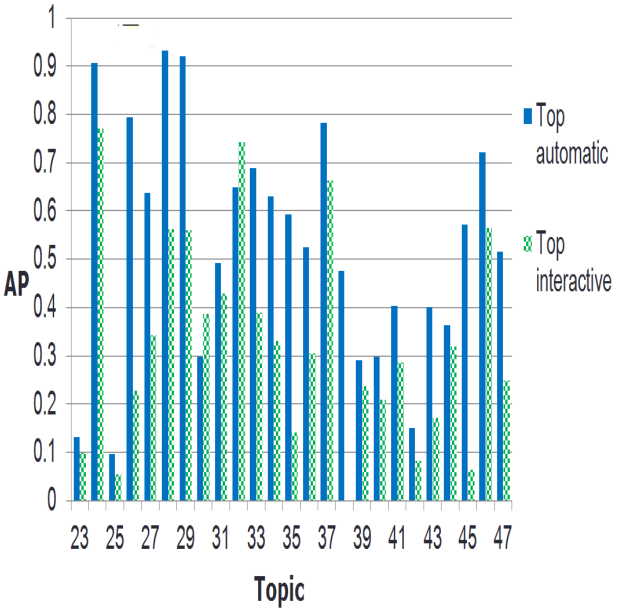


Fig. 21 2012: Average precision for automatic runs by topic/type

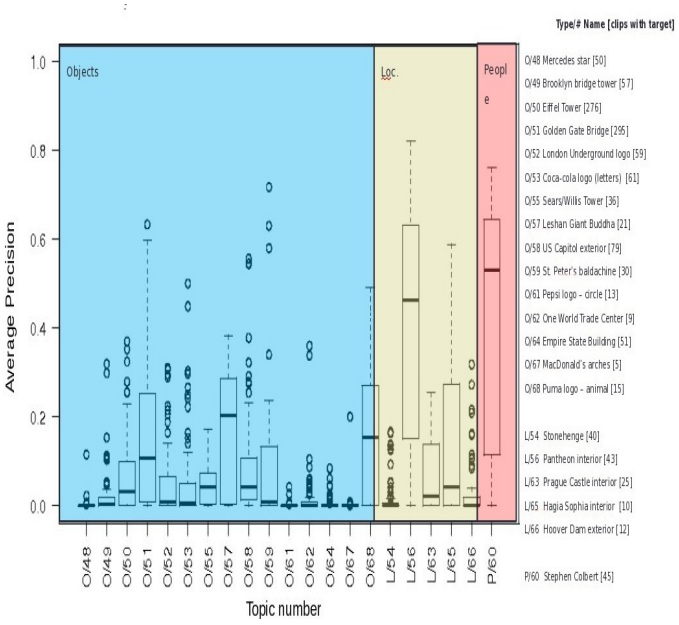


Fig. 22 2012: Average precision for interactive runs by topic/type

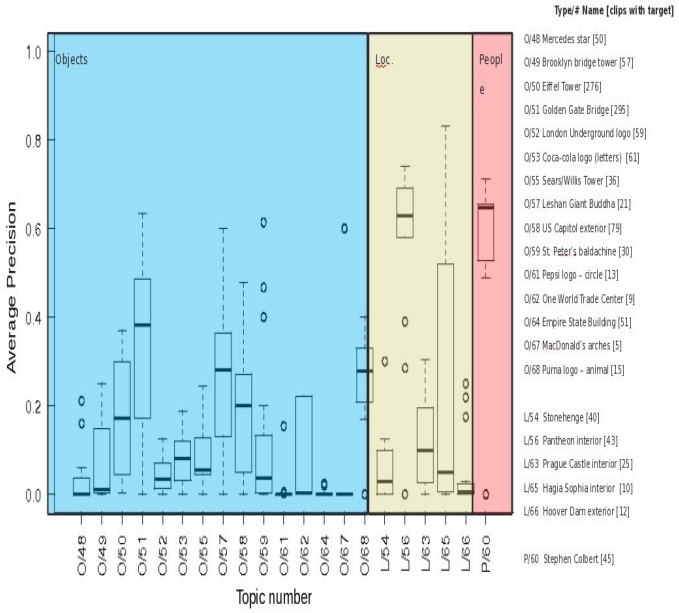


Fig. 23 2012: AP by topic for top runs

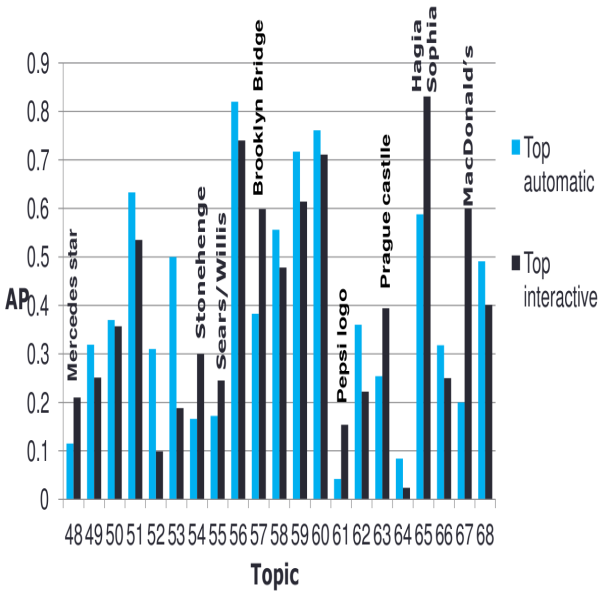


Fig. 24 2013: Boxplot of automatic runs - average precision by topic

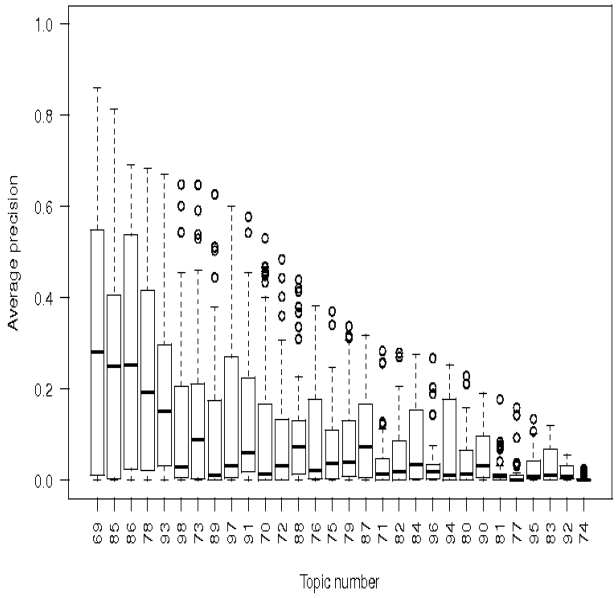


Fig. 25 2013: Boxplot of interactive runs - average precision by topic

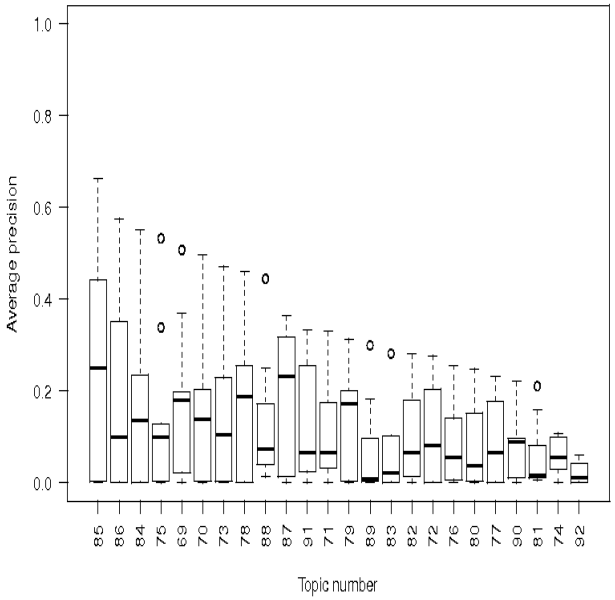


Fig. 26 2014: Boxplot of average precision by topic for automatic runs

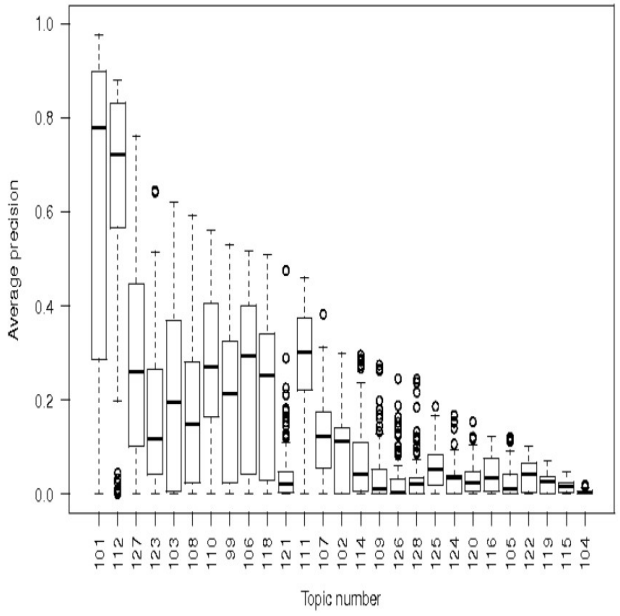


Fig. 27 2014: Boxplot of average precision by topic for interactive runs

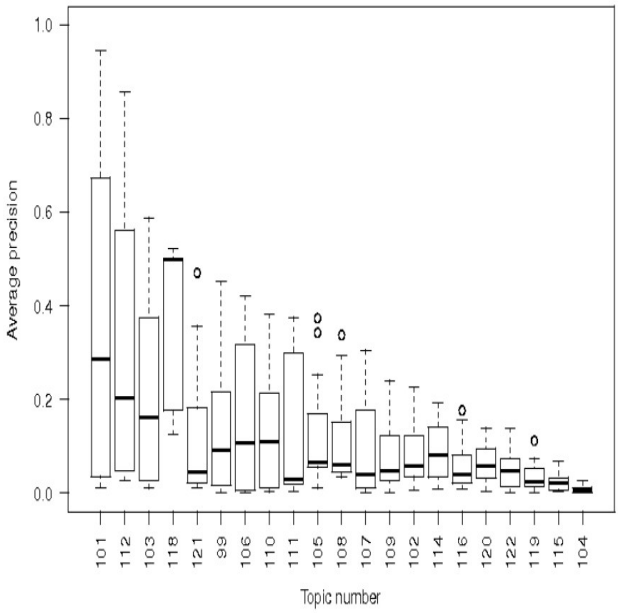


Fig. 28 2014: Effect of number of topic example images used

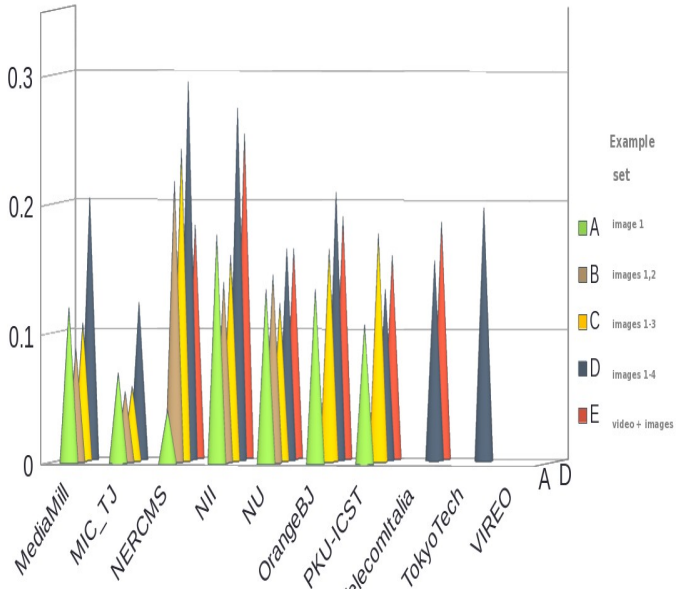


Fig. 29 2015: Boxplot of average precision by topic for automatic runs

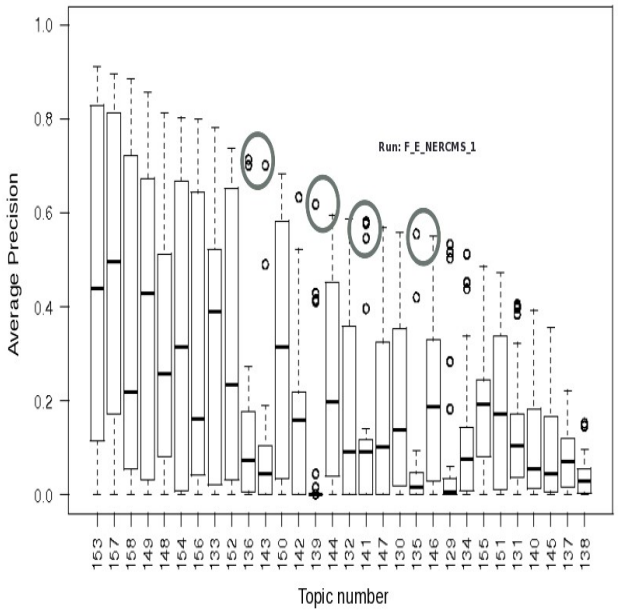


Fig. 30 2015: Boxplot of average precision by topic for interactive runs

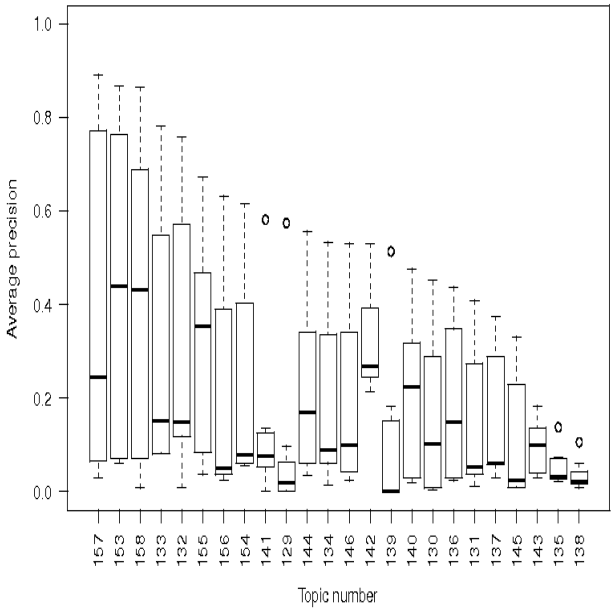


Fig. 31 2015: Automatic results by example Sets (image-only vs video+image)

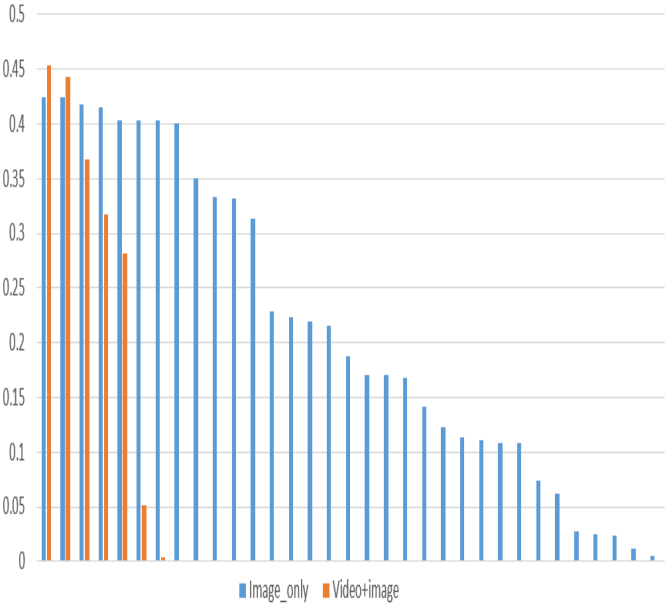


Fig. 32 2013-2015: Mobile vs. Stationary

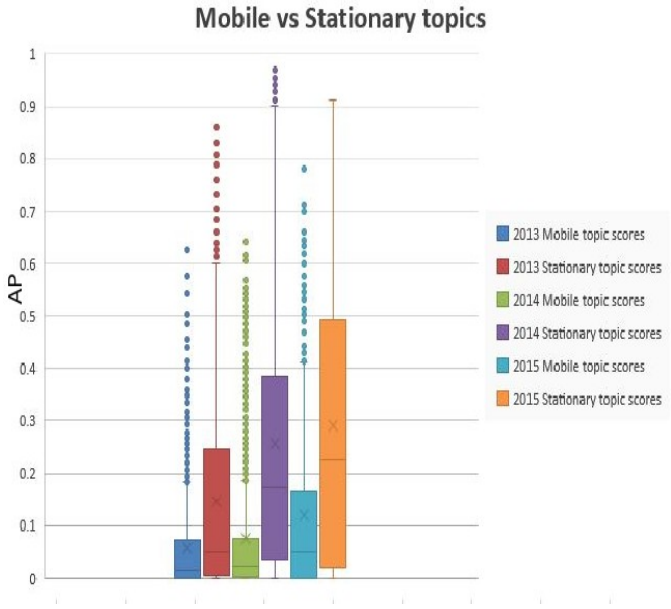


Fig. 33 2013: Box plots of score distributions per topic type (A,B,C,D)-(M,S) pair

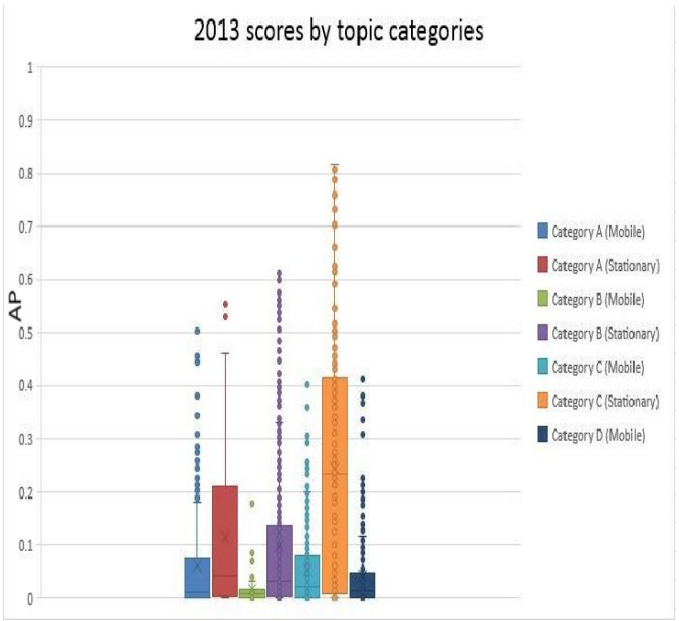


Fig. 34 2014: Box plots of score distributions per topic type (A,B,C,D)-(M,S) pair

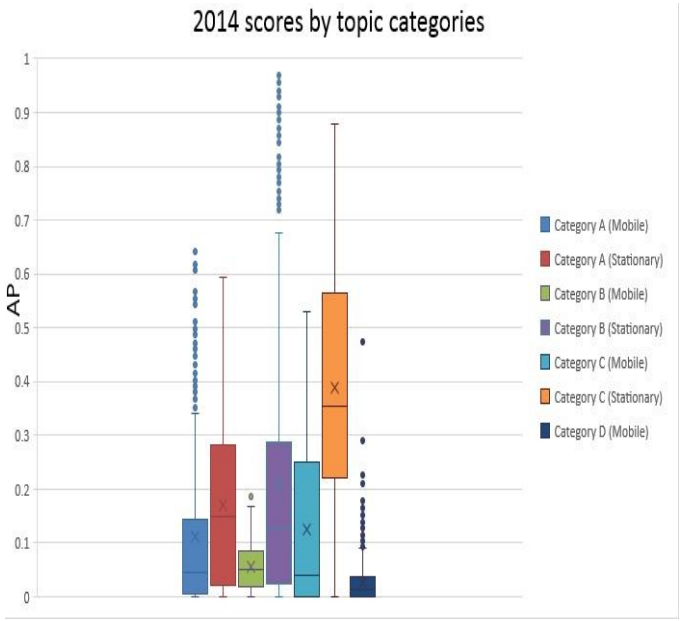
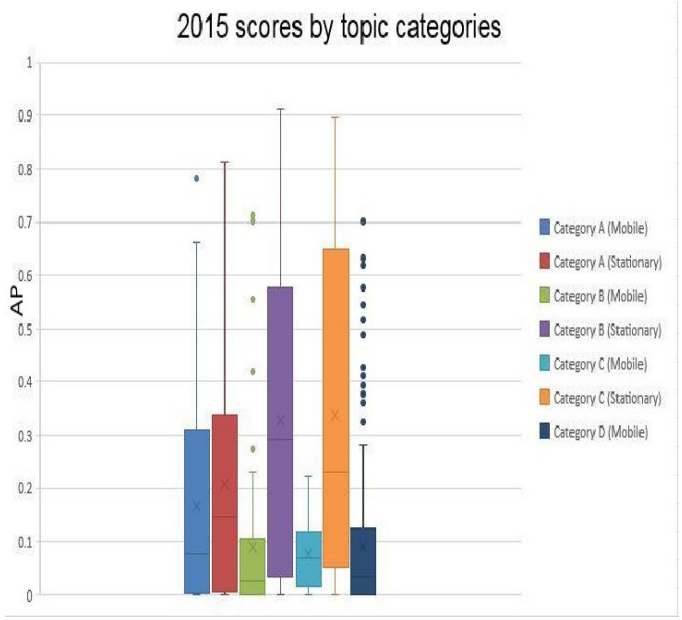


Fig. 35 2015: Box plots of score distributions per topic type (A,B,C,D)-(M,S) pair



6 Overview of TRECVID approaches (2010-2016)

In the previous section, we presented the results of the three pilot years and the three Eastenders years and focused our discussion on the development of the benchmark task including topic creation, overall results in

Fig. 36 Examples of queries with lowest median AP scores

mean average precision and processing time. In this section we summarize some of the main experiments conducted by the participants and relate them to developments in the computer vision and multimedia information retrieval literature as presented in section 2. We omit a discussion of approaches of TV2010 (the first pilot year), since we consider these results not reliable enough to draw meaningful conclusions. [1]

6.1 Summary of approaches at TV2011

The TV2011 INS evaluation displayed a rich set of contrastive experiments performed by the individual teams. Many teams experimented with variants of local features-based representation, combining these, quantizing or not, experimenting with ROI-based filtering and multiple sample images. Some teams tried to enhance results by adding face detection in the processing pipeline. In general, straightforward SIFT (or SIFT variant) based runs achieved the most competitive results.

Retrieval effectiveness: Best results on the TV2011 INS dataset in terms of retrieval effectiveness (MAP=0.531) were achieved by a NII (National Institute of Informatics, Japan) system building on the proven paths of sparse local SIFT descriptors, quantized into a 1M vocabulary to reduce the dimensionality.

Each clip was represented by a single histogram, possibly weighted by an idf component. Ranking was performed by histogram matching (in one of the variants rather similar to tf-idf weighting) resulting in a classical BoVW approach. Advantage was taken of the mask image, for dense sampling local points to boost performance for small instances. This system took about 15 minutes online processing time for each topic. Another strong system (BUPT: Beijing University of Posts and Telecommunications, MAP = 0.407) combined 9 different types of features (global, regional and local) with an elaborate fusion strategy. The system performed well, the small size of the BoW dictionary size (1K) probably being compensated by the aggregation of different feature types.

Search Efficiency: It is a hard trade-off to combine strong effectiveness with efficient search. The most effective run from NII team took about 15 s processing time, while the next best run with MAP 0.407 from BUPT team was able to acheive processing time of 40 s.

Other approaches: TNO (the Netherlands Organization for Applied Scientific Research) submitted 3 runs. One used an exhaustive keypoint search, one a bag-of-visual-words approach, and one open-source face recognition software. In terms of effectiveness, it was found that the keypoint search significantly outperformed the bag-of-visualwords approach and that face-recognition software can contribute if the queries contain large frontal faces. The NII team explored three different approaches: a) large vocabulary quantization by hierarchical k-means and a weighted histogram intersection based ranking metric, b) combination of similarities based on Global quantization of two sets of scale-invariant feature transforms (SIFTs) and color histograms from the full frames, and c) keypoint matching used to compute the similarity between images of the query and images of all videos. AT&T Labs Research based their instance search system on their content-based copy detection work. A baseline run included speeded up robust features (SURF). A normalization technique promoted matches from each query sample image to near the top. They performed outlier analysis, finding weak performance for homogeneous visual characteristics (low contrast, few edges). They experimented with and identified the use of visual content features as a major challenge. BUPT-MCPRL used features such as hue-saturation-value (HSV) histograms, red-green-blue (RGB) moment, SIFT, SURF, CSIFT, Gabor Wavelet, Edge histograms, local binary patterns (LBP), and histograms of oriented gradients (HoG). Higher weight was given for reranking closeup shots. Specific normalization techniques were developed for each modality. Runs were constructed to compare three (non-specified) score merging strategies.

[1] The more elaborate descriptions of individual teams can be found in the notebook papers at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html. Space constraints preclude including bibliographical references for all INS papers for the period 2010 to 2015.

The VIREO: City University of Hong Kong team looked at key differences with search task and content-based copy detection(CCD): region of interest specification, wider definition of relevance than visual copies (e.g., person), and multiple examples with varying conditions (unlike CCD). Their approach incorporated SIFT, BoW, and one keyframe per shot. Their four runs contrasted the following: full matching (vireo b) versus partial matching (vireo m), use of weak geometric information (vireo b) versus stronger spatial configuration (vireo s), and use of face matching (vireo f). There was no clearly winning approach. Performance depended on aspects such as size, context uniformity, etc. Florida International University / University of Miami, in their first participation in the instance search task, employed texture features plus SIFT, Multiple Correspondence Analysis (MCA), and variants enhanced by k-nearest neighbors (KNN) reranking, MCA reranking, SIFT, and 261 extra training images. No significant differences between the runs were found. The Instituto de Matematica e Estatistica, University of Sao Paulo used pyramid histograms of visual words (PHOW) a variant of Dense SIFT (5 pixels distance), and 600 000 descriptors clustered into 300 visual words. Frames were represented as word frequency vectors. The similarity computation was based on chi-square. Only one run was submitted; it scored above median for location topics (where texture was important). The researchers at JRS and Vienna University of Technology fused four different techniques: face detection (Viola Jones) followed by face matching (Gabor wavelets), BoF (bag of features) with codebook size 100, mean shift segments (color segmentation), and SIFT. Fusion took the best result across all topic sample images for all four methods. SIFT-only run performed best, especially well for location type. IRIM team was a large collaboration of European research groups. They used two representations: bag of visual words (BoVW) (using SURF descriptors) 16 000-word codebook and bag of regions (with HSV histogram as descriptor) 2000-word codebook. For measuring similarity they used BoVW (complement of histogram intersection) and bag-of-regions (BOR) (L1-distance). They made limited use of the mask (only over 8 points for BoVW). The best results came from the merged BOVW / BOR and complete frame approaches.

Interactive task: AXES-DCU was the single participant in the interactive task (human-in-the-loop). 30 media students and archive professionals participated in the study. The AXES-DCU system used a pyramid histogram of visual words based on a dense grid of SIFT features at multiple resolutions. Ranking was achieved using a non-linear chi-square SVM. The submitted runs differed solely on the presumed operating point of the searchers (either recall or precision oriented).

6.2 Summary of approaches at TV2012

The TV2012 INS experiments built on the successful strategies of TV2011. All teams used local descriptors, most often quantized into a bag of visual words reduced space. General trends in team experiments were: how to leverage the topic information (multiple images, ROI mask), combinations of features, how to improve BOVW approaches by exploiting spatial constraints

Retrieval effectiveness: Best results on the TV2012 INS dataset in terms of retrieval effectiveness were achieved by BUPT. The former achieved a MAP=0.268 score. The BUPT system was based on the TV2011 entry but with larger BoVW dictionaries (50K and 10K), speed improvements (approximate K-means instead of K-means) and a query expansion strategy, where the top 10 of the initial search results were used as input for individual subsequent queries and result lists are fused according to a heuristically defined exponentially diminishing weighting scheme. Peking University used a similar strategy for their system (fusing multiple global and local keypoint based representations). In addition, their system applied spatial verification techniques, reranking the top ranks with a semi-supervised algorithm - basically pushing down outlier images - and query expansion using Flickr as an external resource given the topic label.

Search Efficiency: In TV2012, the most effective system was also the most efficient system (BUPT) with a search time under one minute per topic. Most probably, the approximate K-means matching strategy played a decisive role. Another fast system (0.16 s) with max MAP of 0.202 was submitted from the VIREO team where they tested different ways to exploit spatial information through comparing the weak geometric consistency checking (WGC) and spatial topology consistency checking using Delaunay Triangulation (DT) based matching [76].

Other approaches: A large variety of exploratory experiments with different objectives were carried out. The main team experiments can be grouped by a number of themes. Systems reused techniques from information retrieval such as dimension reduction using visual words (1k-1M), inverted files for fast lookup, feature weighting (e.g., BM25, tf-idf, RSJ weights as done by NTT-NII team (NTT: Nippon Telegraph and Telephone)), and pseudo-relevance feedback by BUPT-MCPRL.

In terms of system architecture, some teams built an Ad-hoc search system to pre-index all clips in the collection-defined feature space and analyze queries in this space to rank the clips using all local features, BOVW or SOM. On the other hand, other teams built run-time query specific classifiers by analyzing the query to collect external data and define query specific feature space to rank clips accordingly using local features for sample images and/or re-rank with internet sampled images based classifier.

In terms of how to use the query samples, UvA team reported that fusing the focus of the mask region with the background helps while VIREO reported that background context modeling helps as well. In their approach, to diminish the influence of the visual context of a target of interest they applied blurring. Teams AXES and PKU_ICST (PKU: Peking University) collected extra sample images from internet sources. Enlarging query samples did not increase performance as reported by teams JRS and TNO. In general, participants found that fusion of a whole frame run and a masked region of interest run increased performance.

In regard to feature types and representation, CEA compared BOVW with HSV histograms while University of Sheffield compared PHOW features to SIFT. Different fusion experiments were done as well. BUPT-MCPRL run fusion experiments using HSV histograms, RGB moments, SIFT, C-SIFT, Gabor, EDH, LBP, PHOG and HOG features. IRIM reported no significant difference between different fusion strategies for labs features. JRS reported fusion of SIFT and C-SIFT runs did not help while university of Sheffield experimented with fusion using different distance metrics such as Battacharya, Eucledian, and tf-idf.

Another set of experiments were reported dealing with spatial constraints. The spacial information are dropped when local descriptors are used. However, some postfiltering techniques - mostly with encouraging results - were tested. Mediamill reported that spatial filtering helped 7 topics but hurt the others. DFKI used Hough refinement by checking the scale and orientation of matched descriptors. Picsom used pairwise matching of local descriptors which helped their performance. PKU_ICST re-ranked matching keypoints by clustering top results and weeding out outliers, and VIREO team used standard WGC.

Interactive task: This year two teams participated in the interactive task: Axes and PKU. Axes compared different interfaces (tabbed versus untabbed) and interactive feedback versus no feedback. It was found that a tabbed interface was more effective and that user informed feedback consistently improved performance. PKU used the interaction for labeling 25 clips to train an SVM for reranking, which substantially improved retrieval effectiveness.

6.3 Summary of approaches at TV2013

In terms of new ideas, in TV2013 some sites explored ways of leveraging external sample images, some sites experimented with new system architectures (i.e., GPUs or map/reduce distributed processing). In addition, the adaptation of classical IR models for text such as BM25 inspired several groups.

Retrieval effectiveness: Best results on the TV2013 INS dataset in terms of retrieval effectiveness were achieved by the NII team (MAP=0.31). Their system is based on their TV2011 BoVW architecture with a new asymmetric dissimilarity function [83]. NTT achieved a comparable performance with a system with quite similar preprocessing, however with a different ranking procedure based on exponential BM25 [47] with an adapted IDF component. It was also noted that boosting keypoints within the ROI mask was crucial for the TV2013 topics.

Search Efficiency: In TV2013, the most effective systems were also quite efficient. The NII system reported search time at one minute per topic (note that the TV2013 database is at least 6 times larger than TV2012). An even faster system was the VIREO system (6 s per topic) with a decent MAP=0.2 effectiveness score. This system is rather similar to their 2012 entry.

Other approaches: Issues explored included how to exploit the focus versus background of the topic example images (University of Amsterdam, VIREO), the effect of adding extra sample images from Internet sources (AXES:Access to Multimedia), and different levels of fusion, combining different feature types (local, global) (CEA, University of Sheffield, BUPT), Vlad quantization (AXES, ITI-CERTH:Informatics and Telematcis Institute Greece), combining multiple keypoint detectors and multiple descriptors (NII, NTT). The AXES team experimented with finding additional faces using Google image search to enhance the training data. Orange Labs Beijing incorporated a face classifier which helped with some topics at a cost for processing time. Various groups experimented with system architectures and efficiency. TNO used Hadoop to speed up their searches. JRS employed a graphic processing unit (GPU) for object search. The Multimedia and Intelligent Computing Lab at Tongji University team implemented hybrid parallelization using GPUs and map/reduce. A

number of systems incorporated techniques from text information retrieval including inverted files for fast lookup, use of collection statistics (BM25 weighting enhancements NTT-NII), and pseudo-relevance feedback (PKU, NTT-NII, IAD-DCU:Dublin City University.

Interactive task: Interactive experiments were carried out by several teams. For the Orange Labs Beijing team and the PKU team interactive runs outperformed their automatic runs (due to multiple feedback cycles). The AXES group looked at fusion of query-time subsystems (closed captions, Google image visual model, face recognition, object/location retrieval) and their experiments focused on different user types. Three interactive runs from ITI-CERTH found Vlad quantization outperformed BoVW and that their user interface benefited from a scene segmentation module that linked related shots.

6.4 Summary of approaches at TV2014

As usual, nearly all systems used some form of SIFT local descriptors, but there was a large variety of experiments addressing representation, fusion, or efficiency challenges. The trend was moving to larger bag of visual words (BoVW) vocabularies, and larger numbers of keyframes. New in 2014 were several experiments with convolutional neural networks (CNN) for intermediate features. There was increased focus on post-processing (e.g., spatial verification, feedback). The effectiveness of new methods was not consistent across teams, so further research is needed.

Retrieval effectiveness: In TV2014, the best performance was again achieved by the NII team with a MAP of 0.325. Main differentiating new element with respect to their TV2013 entry is an improved spatial consistency enforcement method combining RANSAC and DPM scores. Another strong result was achieved by Nagoya University (MAP=0.304). The Nagoya system in fact shared many system design aspects with the NII system. Main improvements were reported to be due to spatial consistency filtering.

Search efficiency: One of the Nagoya runs combined speed with competitive performance, in particular run F_E_NU_2 with 10 s per topic and MAP=0.290. This is a run using the full query clip (type=E). The best performing NII run, was not optimized for speed and reported processing time of 402 645 s per topic.

Other approaches: System developers addressed the issue of dealing with topic information. Teams considered how to exploit the masks (focus versus background). Mediamill compared mask, full, and fused. BUPT assumed the boundary region of mask contained relevant local points. VIREO experimented with background context modeling using a stare model and found it helps. Teams experimented with combining sample images. Several teams used joint average querying to combine samples into a single query.

Some teams tried exploiting the full video clip for query expansion. NII tracked interest points in ROI and found it helpful sometimes, but interlaced video raised issues. OrangeBJ found no gains. Tokyotech tried tracking and warping the mask with a small gain. VIREO found tracking objects in query video helped if video quality is good (often not the case).

Participating researchers worked on finding an optimal representation for the videos. Teams tried processing more frames (IRIM, Nagoya), combining different feature types (local/global), reviewed techniques and their results (IRIM), combined BoVW and CNN (BUPT). Some groups combined multiple keypoint detectors and multiple descriptors. Nagoya found a single descriptor (Hessian Affine and RootSIFT) was almost as good as a combination of 6, yet was more efficient. ORAND used no quantization codebook, kept raw keypoints, and faced a scaling issue. Sheffield compared SIFT, HOG, global features.

Experiments with MPEG-7 features were carried out by TU Chemnitz and TelecomItalia; they seemed reasonable for mid-sized rigid objects. INSIGHTDCU explored the potential of convolutional neural networks (CNN) in promising experiments with a small-scale dataset. The approach seemed to be useful as a representation that could help improve BOVW, but not sufficiently discriminative for primary search keys.

Several teams experimented with how best to match topics to videos. Typically inverted files were used for fast lookup in sparse BoVW space (Lucene). NII used asymmetric similarity function (2013); it was tested by IRIM to no effect, but Nagoya found it helped. VIREO found a new normalization term in the cosine similarity function which helped to increase recall.

Collection statistics were used by some teams - BM25 enhancements for weighting (NTT-NII) helped, as did IDF adjusted for burstiness (INSIGHTDCU). Pseudo relevance feedback and query expansion were explored by NTT-CSL, who used ROI features for reranking and found it promising.

In studies involving post-filtering, NII tested an improved spatial verification method; Nagoya found that spatial verification helped; OrangeBJ used a face de-

tector for filtering hits for topics involving faces but got no improvement; Wuhan University applied a face filter and color filter; TU Chemnitz employed an indoor/outdoor detector based on audio analysis for removing false matches.

In the matter of system architecture and efficiency JRS experimented with compact VLAT signatures; but a particular signature was not sufficiently discriminative; TU Chemnitz tried PostgreSQL on grid platform; MIC TJ (Tongjing Univ) tried hybrid parallelization using CPU's, GPU's and map/reduce; ORAND approximated K-nearest neighbors (KNN) on unquantized local descriptors; Nagoya worked on efficient re-ranking methods (involving spatial verification); and CERTH built a complete index in RAM.

Interactive task: Several teams built interactive systems. OrangeBJ (BUPT and Orangelabs) had strong performance using a "relative rerank method". BUPT MCPRL used an automatic system without Convolutional Neural Networks for a small gain. ORAND propagated labels to similar shots in same scene using a similarity shot graph. INSIGHTDCU found a system using positive images for new queries outperformed one using them for training an SVM. AXES implemented pseudo relevance feedback and an interactive check. TUC MI (Chemnitz) found MPEG-7 color descriptors were not sufficiently discriminative. ITI CERTH tested shots vs scene presentation and found that shot-based presentation yielded better results.

6.5 Summary of approaches at TV2015

As in previous years, nearly all systems used some form of SIFT local descriptors where a large variety of experiments are addressing representation, fusion or efficiency challenges. In contrast to TV2014, most systems also included a CNN (Convolutional Neural Networks) component. The understanding of when CNN can improve retrieval effectiveness is growing. Many experiments included post-processing (spatial verification, feedback) as an additional step after the ranking stage.

Retrieval effectiveness: In TV2015, the best performance was achieved by the PKU team with a MAP of 0.453, exploiting the full query video (type E run). The PKU system combines a traditional keypoint based architecture with CNN features and spatial consistency checking. PKU also reports to use the video transcripts for reranking, but the benefit of this source is not quantified. Another very strong result was achieved by NII in a collaboration with Hitachi. The NII TV2013 system was further enhanced with a query adaptive late fusion step using a convolutional neural network (MAP=0.424).

Search efficiency: A very fast run (under 2 s) with a reasonable MAP of 0.19 effectiveness was recorded by the InsightDCU system. This run was based on a straightforward BoVW architecture with spatial verification.

Other approaches: A summary of all team efforts in order to find an optimal representation includes: Wuhan team reported improvement from processing more frames, the BUPT and PKU-ICST teams combined different feature types (local/global) and fusion of CNN, SIFT BOW (Bag Of Words) and text captions. LAHORE and SHEFFIELD compared 4 different combinations of 4 different local features and 4 matching methods. Trimps team compared BOW based on SIFT with Faster-RCNN features and global deep features, selective search and CNN with LSH (Locality-Sensitive Hashing) and HOGgles with local features. TU Chemnitz team explored the classification of the audio track as in 2014. UMQG team presented a new approach based on object detection and indexing where CNN was used to describe extracted objects from video decomposition and then matching the query image with nearest object in a codebook and quantization framework.

In regard to exploiting the query images and/or videos the Wuhan team manually selected ROI (region of interest) on different query images which helped significantly their system while exploiting the full query video was applied by PKU ICST, NERCMS, Wuhan and Chemnitz teams. Different matching experiments are reported by systems. Typically inverted files for fast lookup in sparse BoVW space and pseudo relevance feedback for query expansion are mentioned in several reports. Other teams experimented with similarity functions. For example BUPT team used query adaptive late fusion while Wuhan team applied Asymmetric query adaptive matching.

Postprocessing the ranked list results also has been investigated by InsightDCU team where weak geometry consistency check for spatial filtering helped to refine results. The NII-HITACHI team applied DPM (deformable part models) and Fast RCNN in their postprocessing experiments. The Wuhan team applied face and color filters with adjacent shot matching and query text expansion. The NTT team used spatial verification methods such as Ensemble of weak geometric relations and Angle free hough voting in 3D camera motion space. Finally the TU Chemnitz team used indoor/outdoor detectors based on audio analysis for removal of false matches in addition to clustering similar shot sequences.

Interactive task: TU Chemnitz fast reviewed 3500 instances which improved on their automatic results. The PKU ICST team used 2 rounds of relevance feedback on the initial run and fused the results with the original run results. ITI-CERTH evaluated a standard BoVW system, a similar system enhanced with a saliency detection algorithm and concluded that fusing both improved performance.

7 Conclusion

The TRECVID instance search task has proved a successful forum for testing query by visual example techniques for video collections. As such it is a unique task that has produced several benchmarking collections. Developing the task was difficult and took several pilot years to finally converge into a setting with a collection of videos, a collection of topics and a set of measures that have enabled the community interested in this particular problem to gradually improve their techniques over the years. The two main problems in evaluating instance search systems in TRECVID were: finding realistic data with sufficient repeated instances and then creating realistic test topics that fit the data. After three pilot years, the instance search task consolidated towards a task centered around the BBC soap opera series "Eastenders". This kind of series offered sufficient repeated instance, which is a requirement for meaningful statistics and being able to rank systems on search effectiveness. Topic targets were selected to exhibit several kinds of variability - inherent (boundedness, size, rigidity, planarity), locale (multiplicity, variability, complexity), and camera view (distance, angle, lighting). Topics were either stationary, where the immediate context could play a decisive role in classification or mobile, where topic surroundings can vary much more and search is therefore more difficult. The size of the EastEnders data set allowed NIST to create 90 topics across several categories. Several hypothesis about topic 'easiness' have been proposed in Section 3. The aggregated search results of TV INS systems in 2013, 2014 and 2015 were used to validate these hypotheses. It was found that indeed stationary topics are easier than mobile. For stationary topics, type C topics (rigid, planar, logo) seem easier than the non planar types A (rigid, non-planar, small) and B (rigid, non-planar, large). For mobile topics, no consistent rank order could be determined across the 2013 to 2015 topic collections. In general, among the different query types, the objects category seems to be the most type that systems are capable of reporting more progress on. On the other hand, persons and locations seem to be more challenging due to the high variation in how people look or how big is the location boundary.

A basic instance search system can achieve reasonable retrieval effectiveness with close to real-time performance on a database of 470 000 video clips. This performance can be improved with perhaps 30 % by making the system more complex i.e. by a more complex video representation or more extensive key-point filtering after ranking.

Although video queries should be richer in terms of information than image queries, a proper processing pipeline of video queries has not sufficiently been studied. Since low-level video processing techniques are still progressing (such as video object tracking challenge [33] and video tubelet [28]), the processing methodology of video queries will improve.

Unlike other TRECVID tasks, the INS task proves less suitable for machine learning approaches such as support vector machines. Convolutional neural nets are making their way as a supplementary technique. Also, deep learning-based methods did not succeed so far in a significant performance boost of the INS task despite their competitive results in other visual tasks such as image classification [34], face recognition [68], and object detection [22]. Applying deep-learning approaches to enhance instance search is therefore still an open challenge. Apart from the applied machine learning methods, there is still open challenges that systems have to address such as scalability, efficiency, result ranking and similarity measures [78,82].

The TRECVID Instance search task is still ongoing. In fact, since 2016 TRECVID started to explore more complex queries such as retrieving specific target persons at specific locations using the same BBC Eastenders data. A new data collection which goes beyond the soap opera setting would probably extend the impact of the task. In addition, other complex queries such as specific person appearing with specific objects, or perhaps doing specific actions are all interesting scenarios that can test systems capability on not just retrieving instances but combining them.

Finally, all previous years' queries, ground truth and results are available from the TRECVID website and can be accessible from the past data page [70] and yearly proceedings overview slides [71] and participant submitted notebook papers.

Acknowledgements The TRECVID organizers would like to thank Noel O'Connor and Kevin McGuinness at Dublin City University along with Robin Aly at the University of Twente who worked with NIST along with Andy O'Dwyer and William Hayes at the BBC to make the BBC EastEnders video available for use in TRECVID (all Programme material copyrighted by BBC).

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. Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 860–867 (2005). DOI 10.1109/CVPR.2005.81
2. Arandjelovic, R., Zisserman, A.: Multiple queries for large scale specific object retrieval. In: R. Bowden, J.P. Collomosse, K. Mikolajczyk (eds.) British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012, pp. 1–11. BMVA Press (2012). DOI 10.5244/C.26.92. URL http://dx.doi.org/10.5244/C.26.92
3. Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, pp. 2911–2918. IEEE Computer Society (2012). DOI 10.1109/CVPR.2012.6248018. URL http://dx.doi.org/10.1109/CVPR.2012.6248018
4. Babenko, A., Lempitsky, V.S.: Aggregating local deep features for image retrieval. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 1269–1277. IEEE Computer Society (2015). DOI 10.1109/ICCV.2015.150. URL http://dx.doi.org/10.1109/ICCV.2015.150
5. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.S.: Neural codes for image retrieval. In: D.J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, *Lecture Notes in Computer Science*, vol. 8689, pp. 584–599. Springer (2014). DOI 10.1007/978-3-319-10590-1_38. URL http://dx.doi.org/10.1007/978-3-319-10590-1_38
6. Bay, H., Ess, A., Tuytelaars, T., Gool, L.J.V.: Speeded-up robust features (SURF). Computer Vision and Image Understanding **110**(3), 346–359 (2008). DOI 10.1016/j.cviu.2007.09.014. URL http://dx.doi.org/10.1016/j.cviu.2007.09.014
7. Blog, G.M.: Open your eyes: Google Goggles now available on iPhone in Google Mobile App. http://googlemobile.blogspot.com/2010/10/open-your-eyes-google-goggles-now.html (2010)
8. Broder, A.: On the resemblance and containment of documents. In: Compression and Complexity of Sequences 1997. Proceedings, pp. 21–29 (1997). DOI 10.1109/SEQUEN.1997.666900
9. Chandrasekhar, V., Chen, D.M., Tsai, S.S., Cheung, N., Chen, H., Takacs, G., Reznik, Y.A., Vedantham, R., Grzeszczuk, R., Bach, J., Girod, B.: The stanford mobile visual search data set. In: A.C. Begen, K. Mayer-Patel (eds.) Proceedings of the Second Annual ACM SIGMM Conference on Multimedia Systems, MMSys 2011, Santa Clara, CA, USA, February 23-25, 2011, pp. 117–122. ACM (2011). DOI 10.1145/1943552.1943568. URL http://doi.acm.org/10.1145/1943552.1943568
10. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: Finding a (thick) needle in a haystack. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 17–24. IEEE Computer Society (2009). DOI 10.1109/CVPRW.2009.5206531. URL http://dx.doi.org/10.1109/CVPRW.2009.5206531
11. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, pp. 1–8. IEEE (2007). DOI 10.1109/ICCV.2007.4408891. URL http://dx.doi.org/10.1109/ICCV.2007.4408891
12. Chum, O., Philbin, J., Zisserman, A.: Near duplicate image detection: min-hash and tf-idf weighting. In: M. Everingham, C.J. Needham, R. Fraile (eds.) Proceedings of the British Machine Vision Conference 2008, Leeds, September 2008, pp. 1–10. British Machine Vision Association (2008). DOI 10.5244/C.22.50. URL http://dx.doi.org/10.5244/C.22.50
13. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. Workshop on statistical learning in computer vision, ECCV **1**(1-22), 1–2 (2004)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, pp. 886–893. IEEE Computer Society (2005). DOI 10.1109/CVPR.2005.177. URL http://dx.doi.org/10.1109/CVPR.2005.177
15. Davis, W.R., Kosicki, B.B., Boroson, D.M., Kostishack, D.: Micro air vehicles for optical surveillance. Lincoln Laboratory Journal **9**(2), 197–214 (1996)
16. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255. IEEE Computer Society (2009). DOI 10.1109/CVPRW.2009.5206848. URL http://dx.doi.org/10.1109/CVPRW.2009.5206848
17. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015). DOI 10.1007/s11263-014-0733-5. URL http://dx.doi.org/10.1007/s11263-014-0733-5
18. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In: Proceedings of the British Machine Vision Conference (2006)
19. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010). DOI 10.1109/TPAMI.2009.167. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.167
20. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981). DOI 10.1145/358669.358692. URL http://doi.acm.org/10.1145/358669.358692

21. van Gemert, J., Geusebroek, J., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: D.A. Forsyth, P.H.S. Torr, A. Zisserman (eds.) Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III, *Lecture Notes in Computer Science*, vol. 5304, pp. 696–709. Springer (2008). DOI 10.1007/978-3-540-88690-7_52. URL http://dx.doi.org/10.1007/978-3-540-88690-7_52
22. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pp. 580–587. IEEE Computer Society (2014). DOI 10.1109/CVPR.2014.81. URL http://dx.doi.org/10.1109/CVPR.2014.81
23. Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: Multi-pie. Image Vision Comput. **28**(5), 807–813 (2010). DOI 10.1016/j.imavis.2009.08.002. URL http://dx.doi.org/10.1016/j.imavis.2009.08.002
24. Harris, C., Stephens, M.: A combined corner and edge detector. In: C.J. Taylor (ed.) Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988, pp. 1–6. Alvey Vision Club (1988). DOI 10.5244/C.2.23. URL http://dx.doi.org/10.5244/C.2.23
25. Hauptmann, A., Smith, M.: Text, speech, and vision for video segmentation: The informedia tm project. In: Proceeding of AAAI Fall Symposium Computational Models for Integrating Language and Vision, Boston. Citeseer (1995)
26. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: ICCV (2007)
27. Huang, J., Kumar, R., Mitra, M., Zhu, W., Zabih, R.: Image indexing using color correlograms. In: 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico, pp. 762–768. IEEE Computer Society (1997). DOI 10.1109/CVPR.1997.609412. URL http://dx.doi.org/10.1109/CVPR.1997.609412
28. Jain, M., van Gemert, J.C., Jégou, H., Bouthemy, P., Snoek, C.G.M.: Action localization with tubelets from motion. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pp. 740–747. IEEE Computer Society (2014). DOI 10.1109/CVPR.2014.100. URL http://dx.doi.org/10.1109/CVPR.2014.100
29. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: D.A. Forsyth, P.H.S. Torr, A. Zisserman (eds.) Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I, *Lecture Notes in Computer Science*, vol. 5302, pp. 304–317. Springer (2008). DOI 10.1007/978-3-540-88682-2_24. URL http://dx.doi.org/10.1007/978-3-540-88682-2_24
30. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2011). DOI 10.1109/TPAMI.2010.57. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.57
31. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1704–1716 (2012). DOI 10.1109/TPAMI.2011.235. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.235
32. Joly, A., Buisson, O.: Logo retrieval with a contrario visual query expansion. In: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, pp. 581–584. ACM, New York, NY, USA (2009). DOI 10.1145/1631272.1631361. URL http://doi.acm.org/10.1145/1631272.1631361
33. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R.P., Fernández, G., Nebehay, G., Porikli, F., Cehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Trans. Pattern Anal. Mach. Intell. (2016)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., pp. 1106–1114 (2012). URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
35. Le, Q.V., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012. icml.cc / Omnipress (2012). URL http://icml.cc/discuss/2012/73.html
36. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized RANSAC. In: R. Bowden, J.P. Collomosse, K. Mikolajczyk (eds.) British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012, pp. 1–11. BMVA Press (2012). DOI 10.5244/C.26.95. URL http://dx.doi.org/10.5244/C.26.95
37. Letessier, P., Buisson, O., Joly, A.: Scalable mining of small visual objects. In: N. Babaguchi, K. Aizawa, J.R. Smith, S. Satoh, T. Plagemann, X. Hua, R. Yan (eds.) Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012, pp. 599–608. ACM (2012). DOI 10.1145/2393347.2393431. URL http://doi.acm.org/10.1145/2393347.2393431
38. Liu, Z., Li, H., Zhou, W., Tian, Q.: Embedding spatial context information into inverted file for large-scale image retrieval. In: N. Babaguchi, K. Aizawa, J.R. Smith, S. Satoh, T. Plagemann, X. Hua, R. Yan (eds.) Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012, pp. 199–208. ACM (2012). DOI 10.1145/2393347.2393380. URL http://doi.acm.org/10.1145/2393347.2393380
39. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2), 129–136 (1982). DOI 10.1109/TIT.1982.1056489. URL http://dx.doi.org/10.1109/TIT.1982.1056489
40. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004). DOI 10.1023/B:VISI.0000029664.99615.94. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94
41. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vision Comput. **22**(10), 761–767 (2004). DOI 10.1016/j.imavis.2004.02.006. URL http://dx.doi.org/10.1016/j.imavis.2004.02.006
42. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision **60**(1), 63–86 (2004).

DOI 10.1023/B:VISI.0000027790.02288.f2. URL http://dx.doi.org/10.1023/B:VISI.0000027790.02288.f2
43. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
44. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.J.V.: A comparison of affine region detectors. International Journal of Computer Vision **65**(1-2), 43–72 (2005). DOI 10.1007/s11263-005-3848-x. URL http://dx.doi.org/10.1007/s11263-005-3848-x
45. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2227–2240 (2014)
46. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. International Journal of Computer Vision **14**(1), 5–24 (1995). DOI 10.1007/BF01421486. URL http://dx.doi.org/10.1007/BF01421486
47. Murata, M., Nagano, H., Mukai, R., Kashino, K., Satoh, S.: BM25 With Exponential IDF for Instance Search. Multimedia, IEEE Transactions on **16**(6), 1690–1699 (2014). DOI 10.1109/TMM.2014.2323945
48. Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G.: The QBIC project: Querying images by content, using color, texture, and shape. In: Storage and Retrieval for Image and Video Databases (SPIE), pp. 173–187 (1993)
49. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, pp. 2161–2168. IEEE Computer Society (2006). DOI 10.1109/CVPR.2006.264. URL http://dx.doi.org/10.1109/CVPR.2006.264
50. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: A. Leonardis, H. Bischof, A. Pinz (eds.) Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV, *Lecture Notes in Computer Science*, vol. 3954, pp. 490–503. Springer (2006). DOI 10.1007/11744085_38. URL http://dx.doi.org/10.1007/11744085_38
51. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996). DOI 10.1016/0031-3203(95)00067-4. URL http://dx.doi.org/10.1016/0031-3203(95)00067-4
52. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**(3), 145–175 (2001). DOI 10.1023/A:1011139631724. URL http://dx.doi.org/10.1023/A:1011139631724
53. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: K. Daniilidis, P. Maragos, N. Paragios (eds.) Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV, *Lecture Notes in Computer Science*, vol. 6314, pp. 143–156. Springer (2010). DOI 10.1007/978-3-642-15561-1_11. URL http://dx.doi.org/10.1007/978-3-642-15561-1_11
54. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. IEEE Computer Society (2007). DOI 10.1109/CVPR.2007.383172. URL http://dx.doi.org/10.1109/CVPR.2007.383172
55. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society (2008). DOI 10.1109/CVPR.2008.4587635. URL http://dx.doi.org/10.1109/CVPR.2008.4587635
56. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000). DOI 10.1109/34.879790. URL http://doi.ieeecomputersociety.org/10.1109/34.879790
57. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, pp. 1–8. IEEE (2007). DOI 10.1109/ICCV.2007.4408986. URL http://dx.doi.org/10.1109/ICCV.2007.4408986
58. Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications **4**(3), 251–258 (2016)
59. Romberg, S., Lienhart, R.: Bundle min-hashing. IJMIR **2**(4), 243–259 (2013). DOI 10.1007/s13735-013-0040-x. URL http://dx.doi.org/10.1007/s13735-013-0040-x
60. Romberg, S., Pueyo, L.G., Lienhart, R., van Zwol, R.: Scalable logo recognition in real-world images. In: F.G.B.D. Natale, A.D. Bimbo, A. Hanjalic, B.S. Manjunath, S. Satoh (eds.) Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011, p. 25. ACM (2011). DOI 10.1145/1991996.1992021. URL http://doi.acm.org/10.1145/1991996.1992021
61. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2010). DOI 10.1109/TPAMI.2009.154. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154
62. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. International Journal of Computer Vision **37**(2), 151–172 (2000). DOI 10.1023/A:1008199403446. URL http://dx.doi.org/10.1023/A:1008199403446
63. Sivic, J., Everingham, M., Zisserman, A.: Person Spotting: Video Shot Retrieval for Face Sets. In: Proceedings of the ACM International Conference on Image and Video Retrieval (2005). URL http://www.robots.ox.ac.uk/ vgg
64. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 1470–1477 vol.2 (2003). DOI 10.1109/ICCV.2003.1238663
65. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France, pp. 1470–1477. IEEE Computer Society (2003). DOI 10.1109/ICCV.2003.1238663. URL http://doi.ieeecomputersociety.org/10.1109/ICCV.2003.1238663
66. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the

early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000). DOI 10.1109/34.895972. URL http://doi.ieeecomputersociety.org/10.1109/34.895972
67. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1), 11–32 (1991). DOI 10.1007/BF00130487. URL http://dx.doi.org/10.1007/BF00130487
68. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, pp. 1701–1708. IEEE Computer Society, Washington, DC, USA (2014). DOI 10.1109/CVPR.2014.220. URL http://dx.doi.org/10.1109/CVPR.2014.220
69. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. In: Proc. of International Conference on Learning Representations (2016)
70. TRECVID: Past data page. http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html (2001)
71. TRECVID: Trec video retrieval evaluation notebook papers and slides. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html (2001)
72. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: What is the spatial extent of an object? In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 770–777. IEEE Computer Society (2009). DOI 10.1109/CVPRW.2009.5206663. URL http://dx.doi.org/10.1109/CVPRW.2009.5206663
73. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 25–32. IEEE Computer Society (2009). DOI 10.1109/CVPRW.2009.5206566. URL http://dx.doi.org/10.1109/CVPRW.2009.5206566
74. Yang, J., Yu, K., Gong, Y., Huang, T.S.: Linear spatial pyramid matching using sparse coding for image classification. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 1794–1801. IEEE Computer Society (2009). DOI 10.1109/CVPRW.2009.5206757. URL http://dx.doi.org/10.1109/CVPRW.2009.5206757
75. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73**(2), 213–238 (2007). DOI 10.1007/s11263-006-9794-4. URL http://dx.doi.org/10.1007/s11263-006-9794-4
76. Zhang, W., Ngo, C.: Topological spatial verification for instance search. IEEE Transactions on Multimedia **17**(8), 1236–1247 (2015). DOI 10.1109/TMM.2015.2440997. URL http://dx.doi.org/10.1109/TMM.2015.2440997
77. Zhang, Y., Jia, Z., Chen, T.: Image retrieval with geometry-preserving visual phrases. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011, pp. 809–816. IEEE Computer Society (2011). DOI 10.1109/CVPR.2011.5995528. URL http://dx.doi.org/10.1109/CVPR.2011.5995528
78. Zhao, W.L., Ngo, C.W., Wang, H.: Fast covariant vlad for image search. IEEE Transactions on Multimedia **18**(9), 1843–1854 (2016). DOI 10.1109/TMM.2016.2585023
79. Zhu, C., Huang, Y., Satoh, S.: Multi-image aggregation for better visual object retrieval. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pp. 4304–4308. IEEE (2014). DOI 10.1109/ICASSP.2014.6854414. URL http://dx.doi.org/10.1109/ICASSP.2014.6854414
80. Zhu, C., Satoh, S.: Large vocabulary quantization for searching instances from videos. In: H.H. Ip, Y. Rui (eds.) International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012, p. 52. ACM (2012). DOI 10.1145/2324796.2324856. URL http://doi.acm.org/10.1145/2324796.2324856
81. Zhu, C., Satoh, S.: Evaluation of visual object retrieval datasets. In: IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013, pp. 3954–3958. IEEE (2013). DOI 10.1109/ICIP.2013.6738814. URL http://dx.doi.org/10.1109/ICIP.2013.6738814
82. Zhu, C.Z., Jégou, H., Ichi Satoh, S.: Query-adaptive asymmetrical dissimilarities for visual object retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1705–1712 (2013)
83. Zhu, C.Z., Jégou, H., Satoh, S.: Query-adaptive asymmetrical dissimilarities for visual object retrieval. In: The IEEE International Conference on Computer Vision (ICCV) (2013)
84. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comput. Surv. **38**(2) (2006). DOI 10.1145/1132956.1132959. URL http://doi.acm.org/10.1145/1132956.1132959