# Stanford I2V: A News Video Dataset for Query-by-Image Experiments

André Araujo
Stanford University, CA
afaraujo@stanford.edu

Jason Chaves
Stanford University, CA
jchaves@stanford.edu

David Chen
Stanford University, CA
dmchen@stanford.edu

Roland Angst
Stanford University, CA
rangst@stanford.edu

Bernd Girod
Stanford University, CA
bgirod@stanford.edu

## ABSTRACT

Reproducible research in the area of visual search depends on the availability of large annotated datasets. In this paper, we address the problem of querying a video database by images that might share some contents with one or more video clips. We present a new large dataset, called *Stanford I2V*. We have collected more than 3,800 hours of newscast videos and annotated more than 200 ground-truth queries. In the following, the dataset is described in detail, the collection methodology is outlined and retrieval performance for a benchmark algorithm is presented. These results may serve as a baseline for future research and provide an example of the intended use of the Stanford I2V dataset. The dataset can be downloaded at http://purl.stanford.edu/zx935qw7203.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

query-by-image, video dataset, video indexing, video search

## 1. INTRODUCTION

Visual search is the problem of indexing and querying a large collection of visual data. There exist several variations of this problem, depending on the type of content in the database and the type of query. For example, image-to-image (I2I) visual search can be used for product search using an image taken with a mobile device. Video-to-video (V2V) is commonly used for copyright enforcement in online video-sharing websites. Video-to-image (V2I) is useful for augmenting the world seen by a head-mounted camera. Yet another flavour is image-to-video (I2V) visual search, where an image-based query is issued to retrieve relevant videos. Example applications for I2V are advertisement monitoring, video lecture search using slides, organizing and searching a personal video collection or an archive of video, and content linking where a relevant video has to be found based on an image from a certain event (e.g., from a website or news article). The last example is especially challenging since the query image can contain substantial geometric as well as photometric distortion with respect to the frames in the video sequence showing the same event or object. Moreover, in some cases, the query depicts a clean image which is free of any background clutter present in the frames of the video, introducing another asymmetry between query images and frames in the database.

While preparing a database for large-scale I2I search experiments is already challenging, doing the same for I2V search is even more difficult. The handling of video data (ie., acquisition, processing, building and searching data structures) is much more involved, due to the significantly larger data volume. Ground truth annotation with sub-second temporal accuracy can be extremely tedious and time-consuming.

Due to these reasons, previous work was mostly limited to evaluations on small or medium scale benchmark datasets. However, we think that I2V is now at a stage where algorithms and systems have to be evaluated on a larger scale to draw conclusions about their performance. Hence, in this paper, we introduce a new dataset called *Stanford I2V*, consisting of more than 3,800 hours of newscast video and more than 200 queries with ground-truth annotations. We refer to Fig. 1 for an illustration of some queries and relevant video keyframes. The full dataset can be downloaded at http://purl.stanford.edu/zx935qw7203.

In the remainder of this paper, we discuss related work, present Stanford I2V in more detail, outline how the dataset has been collected, and define an evaluation procedure. Furthermore, we present a baseline algorithm which illustrates how the dataset can be used.

## 2. RELATED WORK

Since our main focus is on I2V visual search, we refer the interested reader to recent work [20], the YouTube dataset [14], the HMDB dataset [13] and references therein for more details about V2V (note that these datasets are designed for action recognition rather than retrieval purposes).

Compared to I2I, there is surprisingly little previous work addressing the problem of I2V visual search. One of the early I2V milestones is Sivic and Zisserman's Video Google work [18, 19] where techniques from text retrieval have been


*MMSys'15,* March 18–20, 2015, Portland, OR, USA.

ACM 978-1-4503-3351-1/15/03$15.00
http://dx.doi.org/10.1145/2713168.2713197.

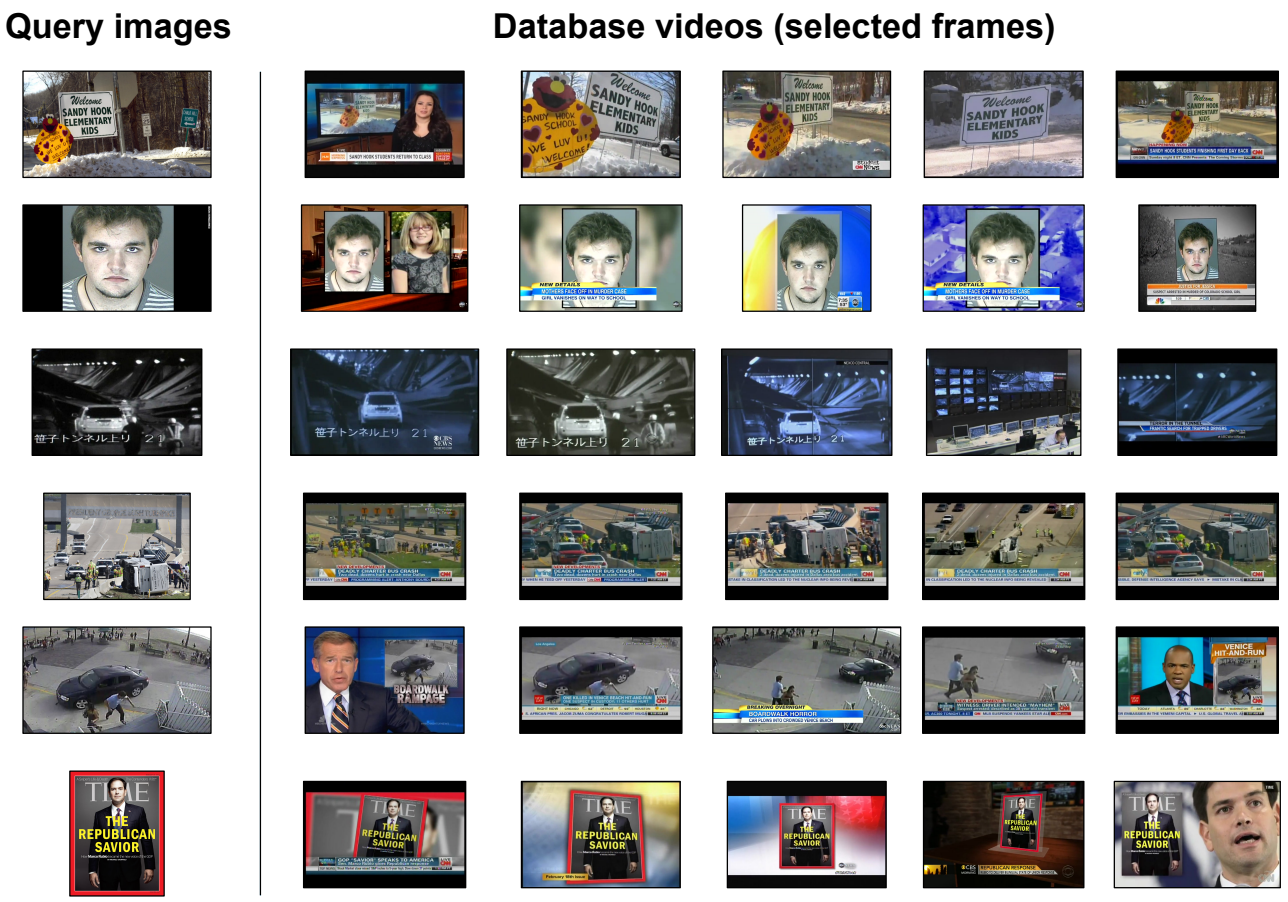


Figure 1: Examples of query images (left) and database video frames (on the right).

| Dataset Name | Availability | Size | Type of Video | # Queries | Type of Query | Unit of Retrieval |
|---|---|---|---|---|---|---|
| Video-Google [18] | no | 2h | Movie: 'Run Lola Run' | 164 | Entire frames from 48 shots taken at 19 different locations | Keyframe |
| Video-Google [18, 19] | no | 2h | Movie: 'Run Lola Run', 'Groundhog Day' | 8 | Small region-of-interest in a frame. | Shot |
| TRECVID-INS13 [16] | yes | 464h | TV soap opera: BBC EastEnders | 30 | Region-of-interest in up to 4 frames | Shot |
| CNN2h [5] | yes | 2h | Newscast: 2 newscasts from CNN | 139 | Images collected from websites and photos taken from a screen | Keyframe |
| Stanford I2V (Ours) | yes | 3,800h | Newscast: 39 recurring newscasts from 25 channels | 229 | Images collected from news websites | Scene and segment |

Table 1: Comparison of I2V datasets: The second column denotes whether the dataset is publicly available. We see that our new Stanford I2V dataset is orders of magnitude larger than any previous dataset. Newscast videos are less constrained than 'closed worlds' such as movies or soap operas and thus expected to be more challenging.

adapted to be applicable for I2V retrieval. They addressed two retrieval scenarios: the first uses an entire frame from a movie as a query to find other frames in the movie which show the same location, whereas the second one uses a small region-of-interest in a frame to select an object of interest which then is detected in the remaining frames.

The dataset released by the TRECVID Instance Search (INS) challenges [16] is most closely related to ours. We refer the interested reader to the TRECVID website[1] for more details about challenges of previous years – here we will focus on the most recent editions of this challenge which used footage from the BBC soap opera EastEnders. A query in that dataset consists of regions-of-interest in up to four frames which denote the outline of an object of interest. While the TRECVID dataset is certainly an improvement over existing datasets, it is still an order of magnitude smaller both in

[1] http://trecvid.nist.gov

database size and number of queries as our new Stanford I2V dataset. Moreover, all the previously described datasets consider a 'closed world' and therefore reflect the variety of real footage only to some extent: the queries and frames in the database are from a movie or soap opera and hence show a limited number of people, objects and locations. The 'closed world' problem is also evidenced by the fact that queries are based on entire frames or regions-of-interest of a frame and therefore use almost exactly the same underlying pixel values as the frames contained in the database. This ignores several important distortions, as mentioned in Sec. 1. Motivated by these shortcomings, we have recently introduced the CNN2h dataset [5] which contains 2 hours of newscast from CNN. Similar to [5], the queries of Stanford I2V are not regions-of-interest of the original frames, but rather images which were collected from news websites that reported about the same event as a certain video. This mimics more realistic use-cases for applications such as the ones mentioned above. The new Stanford I2V dataset is however orders of magnitude larger than the CNN2h (3,800 hours vs. 2 hours). We refer to Tab. 1 for a comparison of previously used I2V datasets.

In summary, previous datasets for I2V (e.g., TRECVID INS [16], CNN2h [5]) are simply not sufficiently diverse or large enough to reveal the entire set of challenges involved in indexing a large collection of videos.

# 3. DATASET DESCRIPTION

The Stanford I2V dataset is a large dataset to evaluate the task of retrieving videos using images as queries. Tab. 2 provides statistics on the dataset composition. The full version of the dataset contains 3.8k hours of video, distributed across 84k video clips on average 2.7 minutes long. The light version of the dataset is a subset of the full version, with the intended use of faster experimentation. Each video clip in our database corresponds to a single news story, segmented from a full-length newscast. These story clips are assembled from a coherent collection of successive shots which cover a single event. Hence, each story clip usually contains tens of shots. These story clips, in the context of news videos, are the equivalent of 'scenes' [21, 22] for general-purpose videos. While for some applications those scenes are the appropriate unit of retrieval, there are other use cases where a more fine-grained unit of retrieval is required. Hence, for our dataset, we tried to strike a good balance and provide annotations at two levels of granularity, namely at the scene level and at the segment level. A segment is a subsequence of a scene with an exact start and end point in which a query contents is visible in the video.

The dataset is accompanied by a carefully selected set of queries with ground truth annotations. Image queries are collected from news websites, and they usually depict important events. The full version of the dataset contains 229 queries. For each query image, we provide a list of all database clips where it is found, along with a list of all precise segments it is shown in the clips.

In order to mimic a broad set of applications, we define the evaluation procedure for this dataset based on two stages, reflecting the two levels of annotation granularity, see also Fig. 2. In the first stage, called *Scene Retrieval*, the objective is to return the correct story clips in the top of a ranked list. The second stage is *Temporal Refinement* where, given a story clip, the precise segments where the query image is visible have to be found.

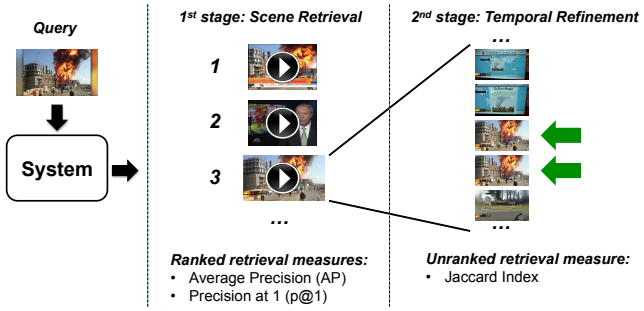


Figure 2: We define the search process based on two steps. First, *Scene Retrieval*: the system returns a ranked list of the most likely story clips to contain the query image. Second, *Temporal Refinement*: if the user is interested in a given clip, the system returns the specific segments within the clip that contain the query image.

# 4. DATASET CONSTRUCTION

## 4.1 Collecting Videos

In previous work [8], we have developed a system that records video newscasts (from cable boxes, over-the-air antennae and video-sharing websites) and segments them into individual stories. Using data resulting from that system, we have demonstrated a real-time system that searches the most recent clips using images [4]. In this work, we consider this scenario at a much larger scale. Our video database is composed of all video clips collected (by the system from [8]) from October 2012 to (and including) September 2013, from 39 different recurring newscasts in 25 different channels. Tab. 2 reports some statistics from this video collection.

## 4.2 Query Set Selection

To find candidate query images for our video database, we used the Internet Archive's Wayback Machine [3]. This tool allows users to browse archived versions of a large number of webpages. For major websites, several webpage captures are stored per day.

Two annotators browsed archived webpages for each day in the date range of October $1^{st}$, 2012 to September $30^{th}$, 2013. The annotators accessed webpages from many different news organizations, usually using Google News [2] as a starting point for each date. 805 candidate query images were collected.

Two types of images were collected: 1) iconic images, i.e., images of events that were reported in the news, and 2) magazine covers from "The Economist" and "Time". For each collected image, we recorded the date it was published online. The collected images are more likely to be shown in news videos that were broadcast around the image's publishing date. During the annotation process, we take advantage of this observation, as explained in the following subsection.

## 4.3 Ground Truth Annotation

The annotation pipeline is illustrated in Fig 3. It contains multiple stages, some of them automatic (in blue) and some of them manual (in orange). Three trained annotators participated in the manual stages. Since the database is composed of videos, we annotate ground-truth video sequences, instead of ground-truth frames. For example, if query image 32 is shown in video clip 21 from 0:38 to 1:34, then this video sequence is one of the ground-truth sequences for query 32.

| | # Video hours | # Queries | # Video clips | # Keyframes @ 1fps | Average clip duration (min.) |
|---|---|---|---|---|---|
| Full version | 3,801 | 229 | 84,443 | 13,966,820 | 2.70 |
| Light version | 1,035 | 78 | 23,437 | 3,808,760 | 2.65 |

Table 2: Statistics of the Stanford I2V dataset. The light version of the dataset is a subset of the full version.

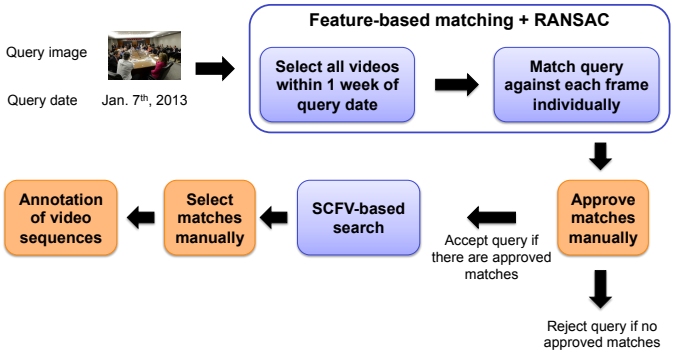


Figure 3: Block diagram of our annotation pipeline. Tasks in blue correspond to automatic stages, while tasks in orange correspond to manual stages. We use frames extracted from videos at 1 fps.

We compute SIFT features [15] (on average, 300 per video frame) and Scalable Compressed Fisher Vector (SCFV) global signatures [9] for each candidate query image and for each frame extracted at 1 fps from the database videos. We exploit the fact that the types of queries we consider are time-sensitive, i.e., they are more likely to have been broadcast in a certain date range. The different steps in the annotation process are described in more detail in the following.

**Feature-based Matching + RANSAC.** In this stage, we perform a fine-grained search of clips that are more likely to contain the candidate query image. All clips within 1 week of the candidate query image date are used. Since the image usually comes from a news piece, after 1 week the particular event is unlikely to be discussed in news videos. The candidate query image is matched against each frame in each clip in the selected time frame. We use the feature-based matching process of [15], using the ratio test. The feature matches are then verified geometrically using RANSAC [10] with an affine model. We declare a match if the discovered geometric model between the query image and the video frame finds at least 10 inliers.

**Approve matches manually.** Once we have candidate matches between query images and video frames, annotators visually inspect them to decide if they are valid. At this stage, we consider a match to be valid if at least part of the query image is shown in the video frame. After this step, if all candidate matches are rejected, then the candidate query is discarded. Otherwise, we continue the process as follows.

**SCFV-based search.** Even though the query image is more likely to be shown in news videos within a certain date range, we still need to make sure it does not appear in the rest of the videos in our database. We search the entire video database using the SCFV global descriptor, which provides a fast way to generate candidate video frames more likely to contain the query. Each video frame in the database is described using the SCFV descriptor with 192 Gaussian mixture components. The SCFV global descriptor is presented in more detail in 5.2.

The 2,000 video frames with smallest global signature distance to the query are further re-ranked using the same feature-based matching and RANSAC scheme as described previously. The final ranked list of candidate frame matches is ranked first by the number of inliers, then by closeness of the SCFV global signature. Since the clips within 1 week of the query publishing date were already considered in the first step, in this step only the remaining clips in the database were analyzed.

**Select matches manually.** The annotators visually inspect the ranked list of video frames resulting from the previous step. Whenever a new correct frame match is found, the corresponding clip is selected and added to the list of ground-truth video clips for the query under consideration.

**Annotation of video sequences.** At this point, a set of video clips are known to contain the query image, from the outcome of the previous steps. One annotator will watch each ground-truth clip and record the exact time sequences where the query image is shown. Note that multiple ground-truth sequences might exist for a ground-truth clip.

For some queries, some potential matches were ambiguous. For example, queries would depict an event from a given angle and the video would show the event from a very different angle, such that the visual similarity would be very small. This scenario was considered in a case-by-case basis by the annotators, and we made sure that the annotations were consistent for all queries. If the annotation for a certain query were too ambiguous, the query was simply discarded.

**Post-processing.** It is still possible that some ground-truth video matches have been missed using the method outlined in this section, since it is not feasible for the annotators to compare all 3,800 hours of video with each query. For this reason, we continue to visually inspect the retrieval results produced by algorithms we experiment with. It has been necessary to add annotations in only a few cases so far. If any additional missing annotations are discovered in the future, we will update the dataset accordingly.

# 5. EVALUATION METRICS

In this section, the performance assessment protocol for the Stanford I2V dataset is presented. We also describe experimental results using a standard image retrieval system, which may serve as a baseline for future research.

## 5.1 Experimental Setup

We divide the experiments for this dataset in two stages. Fig. 2 illustrates an example. First, the system retrieves from the large database those clips that are more likely to contain the query image. This first stage is called *Scene Retrieval*, since each clip in our case corresponds to a scene (a news story). We choose to use story clips as the unit of retrieval in this type of application, since they are concise (on average 2.7 minutes) and meaningful by themselves. Note that related work typically considers retrieval based on shots [16] or frames [5].

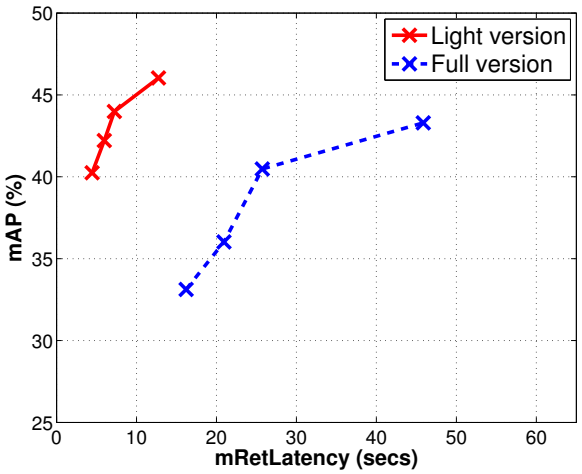


Figure 4: Mean Average Precision (mAP) as a function of mean Scene Retrieval Latency (mRetLatency) for the light and full versions of the Stanford I2V dataset.

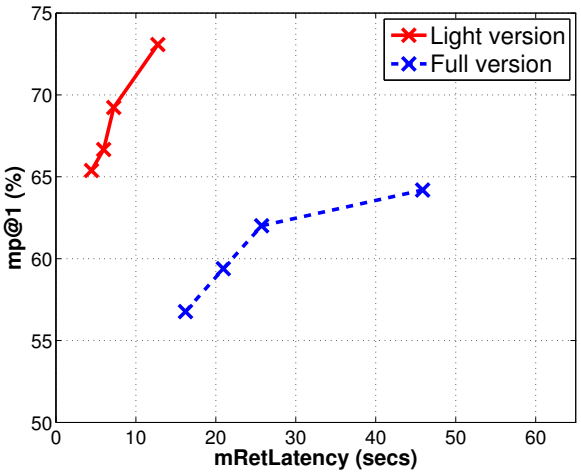


Figure 5: Mean Precision at 1 (mp@1) as a function of mean Scene Retrieval Latency (mRetLatency) for the light and full versions of the Stanford I2V dataset.

The first step is considered a ranked retrieval type of problem, and we measure performance in this case using Average Precision (AP) and Precision at 1 (p@1). Average Precision assesses the quality of the returned ranked list of results and is useful in applications where a list of potential results is shown to the user. Precision at 1 is important in cases where the best result is directly returned to the user (for example, in the case where the system would start playing the best clip match without further interaction with the user).

In the second step, if the user is interested in a particular clip, the system should indicate which points in time in the clip the query image was found. We denote this second step *Temporal Refinement*. In practice, the system could present these matches by showing ticks on the video player. For this second stage, we have an unranked retrieval case, where a match should be presented to the user if the system is confident enough. Since the system may retrieve one or more segments within each ground-truth clip, we assess performance in this case using the Jaccard index. In this case, the Jaccard index is computed by the ratio between the intersection of the retrieved and ground-truth sequences, and their union. This has also been called 'overlap accuracy' in the literature of activity detection [6]. Implementations of scoring functions are provided on the dataset website.

Note that, for Temporal Refinement evaluation, we consider that the correct story clip is given, so the system only needs to find the correct segments within the given clip. Also, in practice, we observed that the precise time segment annotations might vary by up to 1 second due to the usage of different video players. To avoid incorrect scoring, we introduce 1 extra second at the beginning and at the end times of the sequence. For example, if a ground-truth sequence defined by our annotation process starts at 1:12 and ends at 1:23, we will score the retrieved sequence with respect to the time segment starting at 1:11 and ending at 1:24.

In the experiments that follow, we also report results based on query latency for the two steps (using a single core on an Intel Xeon 2.4GHz processor) and total database memory usage. For results over a set of queries, we report mean Average Precision (mAP), mean Precision at 1 (mp@1), mean Jaccard index (mJac), mean scene retrieval latency (mRetLatency) and mean temporal refinement latency (mRefLatency).

## 5.2 Evaluation using standard image retrieval approaches

In our experiments, we use the state-of-the-art global image descriptor SCFV [9], which has been selected by the MPEG Compact Descriptors for Visual Search (CDVS) subgroup for adoption into the CDVS Test Model. The number of Gaussian mixture components has been varied $K \in \{128, 192, 256, 512\}$ to obtain a trade-off in terms of retrieval, memory usage and search latency. In our case, we simply use all Gaussian mixture components, without making use of SCFV's scalability option. We use Fisher Vectors including only the gradients with respect to the mean, which is the most common choice [12, 9]. Local descriptor dimensionality reduction is performed using PCA and keeping the 32 dimensions with largest variance, as in [9]. We use a separate set of images (from the INRIA Holidays [11], Oxford Buildings [17] and Pasadena Houses [1] datasets) to train GMMs, PCA and correlation weights. Note that we implement a simple retrieval algorithm that compares the query's global signature to all frames' global signatures. Speed-up could be obtained by using an inverted index (at the cost of additional memory usage) [9], multi-round scoring [7] or parallelization.

**Scene Retrieval stage.** This SCFV-based scheme is used for the Scene Retrieval stage, where we generate a global signature for each keyframe in our database (extracted at 1 fps, see Tab. 2 for numbers). In this first step, we obtain a ranked list of keyframes, according to the query's most similar signatures in the database. From this list, we obtain the top 100 ranked scenes, where a scene score is defined as the best score among all of its constituent keyframes. For each query, we compute AP and p@1 based on the top 100 retrieved scenes.

**Temporal Refinement stage.** To evaluate the Temporal Refinement stage, we consider each ground-truth clip for each query separately. For each ground-truth clip, we find

| | mAP | mp@1 | mJac | mRet-Latency [s] | mRef-Latency [s] | Index size [GB] |
|---|---|---|---|---|---|---|
| **Light** | | | | | | |
| K = 128 | 0.40 | 0.65 | 0.38 | 4.45 | 0.70 | 1.95 |
| K = 192 | 0.42 | 0.67 | 0.38 | 5.97 | 0.70 | 2.93 |
| K = 256 | 0.44 | 0.69 | 0.38 | 7.23 | 0.69 | 3.90 |
| K = 512 | 0.46 | 0.73 | 0.38 | 12.75 | 0.73 | 7.80 |
| **Full** | | | | | | |
| K = 128 | 0.33 | 0.57 | 0.42 | 16.21 | 0.80 | 7.15 |
| K = 192 | 0.36 | 0.59 | 0.42 | 20.94 | 0.80 | 10.73 |
| K = 256 | 0.40 | 0.62 | 0.42 | 25.73 | 0.78 | 14.30 |
| K = 512 | 0.43 | 0.64 | 0.43 | 45.86 | 0.84 | 28.60 |

Table 3: Quantitative retrieval results for the light and full versions of the dataset: $K$ refers to the number of used Gaussian mixture components.

the 50 most similar frames (in terms of SCFV signatures) and try to find a geometric model between the query image and the video frame using feature matching and RANSAC, as in 4.3. A video frame is returned if at least 8 inliers are obtained.

All results are reported for both light and full versions of the datasets. Fig. 4 and Fig. 5 present Scene Retrieval results: mAP and mp@1 as a function of mRetLatency, respectively. Tab. 3 reports Temporal Refinement results and memory usage for the different SCFV database indexes. In general, a larger number of Gaussian mixture components leads to better mAP and mp@1, at the cost of slower retrieval (higher mRetLatency). Note that mJac and mRefLatency do not vary much with the number of Gaussian mixture components, since the search in this case is restricted to a clip. Scene Retrieval latency and index memory usage increase roughly linearly with the number of Gaussian mixture components.

## 6. SUMMARY

In this work, we introduce Stanford I2V, a new query-by-image video search dataset. Compared to existing datasets, Stanford I2V is more diverse and much larger. We introduce the problem, applications and survey related work. The methodology for data collection and ground-truth annotation is described in detail. We also present experiments using standard image retrieval techniques, serving as a baseline for future evaluations and showcasing the intended use of the Stanford I2V dataset.

## 7. REFERENCES


[1] Computational Vision: Archive, Nov. 2014. http://www.vision.caltech.edu/html-files/archive.html.

[2] Google News, Nov. 2014. http://news.google.com.

[3] Wayback Machine, Nov. 2014. http://archive.org/web.

[4] A. Araujo, D. Chen, P. Vajda, and B. Girod. Real-time Query-by-Image Video Search System. In *Proc. ACM-MM*, 2014.

[5] A. Araujo, M. Makar, V. Chandrasekhar, D. Chen, S. Tsai, H. Chen, R. Angst, and B. Girod. Efficient Video Search Using Image Queries. In *Proc. ICIP*, 2014.

[6] C.-Y. Chen and K. Grauman. Efficient Activity Detection with Max-subgraph Search. In *Proc. CVPR*, 2012.

[7] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod. Residual Enhanced Visual Vector as a Compact Signature for Mobile Visual Search. *Signal Processing*, 93(8):2316–2327, 2013.

[8] M. Daneshi, P. Vajda, D. Chen, S. Tsai, M. Yu, A. Araujo, H. Chen, and B. Girod. EigenNews: Generating and Delivering Personalized News Videos. In *Proc. IEEE BRUREC*, 2013.

[9] L. Duan, J. Lin, J. Chen, T. Huang, and W. Gao. Compact Descriptors for Visual Search. *IEEE Multimedia*, 21(3):30–40, 2014.

[10] M. A. Fischler and R. C. Bolles. Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] H. Jégou, M. Douze, and C. Schmid. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *Proc. ECCV*. Springer, 2008.

[12] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating Local Image Descriptors into Compact Codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1704–16, 2012.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. In *Proc. ICCV*, 2011.

[14] J. Liu, J. Luo, and M. Shah. Recognizing Realistic Actions from Videos "In the Wild". In *Proc. CVPR*, 2009.

[15] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. TRECVID*, 2013.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *Proc. CVPR*, 2007.

[18] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proc. ICCV*, 2003.

[19] J. Sivic and A. Zisserman. Video Google: Efficient Visual Search of Videos. In *Toward Category-Level Object Recognition*, pages 127–144. 2006.

[20] E. H. Taralova, F. De la Torre, and M. Hebert. Motion Words for Videos. In *Proc. ECCV*, 2014.

[21] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of Video by Clustering and Graph Analysis. *Computer Vision and Image Understanding*, 71(1):94–109, 1998.

[22] Y. Zhai and M. Shah. Video Scene Segmentation Using Markov Chain Monte Carlo. *IEEE Transactions on Multimedia*, 8(4):686–697, 2006.