Interactive Video Asset Retrieval using Sketched Queries

Stuart James and John Collomosse
Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
{s.james | j.collomosse}@surrey.ac.uk

ABSTRACT

We present a new algorithm for searching video repositories using free-hand sketches. Our queries express both appearance (color, shape) and motion attributes, as well as semantic properties (object labels) enabling *hybrid* queries to be specified. Unlike existing sketch based video retrieval (SBVR) systems that enable hybrid queries of this form, we do not adopt a model fitting/optimization approach to match at query-time. Rather, we create an efficiently searchable index via a novel space-time descriptor that encapsulates all these properties. The real-time performance yielded by our indexing approach enables interactive refinement of search results within a relevance feedback (RF) framework; a unique contribution to SBVR. We evaluate our system over 700 sports footage clips exhibiting a variety of clutter and motion conditions, demonstrating significant accuracy and speed gains over the state of the art.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Sketch based Retrieval; I.4 [**Image Processing and Computer Vision**]: Feature Measurement—*Feature representation*

General Terms

Computer Vision

Keywords

Sketch based Video Retrieval, Relevance Feedback

1. INTRODUCTION

Efficient and intuitive search tools for large video repositories form an essential component of the modern production pipeline. Searching through rushes footage, or discovering archival assets for production re-use, are common usage scenarios for video assets maintained in digital form. Yet existing asset management systems operate over metadata only, using text-based queries that specify the desired semantic concept (e.g. persons or objects present) in the form of user-annotated tags logged against each video clip. Semantics are both useful and convenient to encode via metadata tags, however the appearance or motion of content within the video clip itself is also useful when specifying a desired video asset. Unfortunately the appearance and motion of objects cannot be expressively conveyed via a few metadata tags and consequently is rarely indexed within existing video search tools.

CVMP'14 13-14 November 2014, London, United Kingdom

http://dx.doi.org/10.1145/2668904.2668940.

On the other hand, pictorial queries such as free-hand sketches offer an intuitive mechanism for expressing appearance and motion information. Sketch based retrieval (SBR) techniques have been applied to large-scale image search [13, 6, 15, 14, 18] (SBIR) and to the search of video collections (SBVR) [7, 9, 16] using sketched color and motion indicators. Since sketch ambiguity and limited user depictive skill are common problems with SBVR systems, combined or '*hybrid*' SBVR systems have been developed that fuse semantics and appearance. Such systems accept sketches annotated with semantic object labels [19], to both disambiguate queries and enhance their expressive power.

This paper presents a novel SBVR system of this 'hybrid' variety, comprising two core technical contributions:

1. **Spatio-temporal Indexing.** Existing hybrid SBVR uses optimization approaches that fit the sketch as a model to each video clip in the database. This per-record optimization is slow (in the order of seconds per record) and scales linearly, at best, with database size. By contrast, our system extracts a digital fingerprint (feature vector) from each video clip representing its content, enabling query comparisons several orders of magnitude faster with the potential to scale sub-linearly.
2. **Relevance Feedback.** Users may iteratively work with our system to refine the returned results by flagging a few 'good' (relevant) or 'poor' (irrelevant) matches. This is essential given the ambiguity inherent in sketch (e. g. shape) when faced with a large database, and given the multiple-modalities (e. g. colour, motion, shape, semantics) present within a hybrid sketched query. Although common in the wider information retrieval literature, and fleetingly explored for SBIR [22, 12], relevance feedback has not been seen in any SBVR system before.

Relevance feedback is enabled through the interactive performance obtained via our novel video descriptor for index-based *hybrid* SBVR. Existing hybrid SBVR approaches treat retrieval as an optimization process in which a model derived from the sketch is fitted to evidence in each video clip in turn [9, 19, 17]. The posterior likelihood of the fit drives the position of each video in the ranked

results. Whilst such approaches are robust, query time is in the order of minutes for a search over hundreds of videos. General purpose visual search algorithms more commonly build a search index; compact and efficiently searchable 'fingerprints' distilled from each media item that can be represented and compared in a high-dimensional feature-space. Despite their power and ubiquity, indexing approaches have not been explored for *hybrid* (i. e. appearance + semantic based) SBVR.

Our index is formed using a novel spatio-temporal descriptor that represents the local shape, color, semantic class and motion parameters of objects within quantized space-time chunks of the video sequence. Additionally, we demonstrate greater flexibility through our ability to incorporate background appearance constraints within our search query. Previous hybrid SBVR systems have enabled search only on the foreground (moving) objects within the scene and so have ignored spatial structure in the background.

We have evaluated our proposed system over a database of 700 sports footage video clips containing 12 categories of subjects of differing appearance moving under various motion conditions (an expanded version of the dataset by Hu et al. [17]). We release this dataset alongside our results and ground-truth as a further contribution [1]. We demonstrate that our new *hybrid* SBVR technique using an indexing approach, when used alone, delivers comparable accuracy to the state of the art (optimization approaches of Hu et al. [17, 19]) but at a fraction of the computational cost. When combined with relevance feedback, our indexing approach comfortably exceeds the current state of the art both in terms of accuracy and speed.

2. RELATED WORK

Sketch based Retrieval (SBR) dates back to the mid-nineties, which yielded sketch based image retrieval (SBIR) techniques based on spatial distributions of shape, color, or texture [2]. Spectral representations for SBIR were later explored by Jacobs et al. [20] using wavelets to match 'finger painted' color blob depictions of query images. VisualSEEk explored graph representations of region topology to improve on earlier blob based systems.

Alternatives to the blob based SBIR approaches using monochrome contour or *line based* techniques gained traction via elastic matching [3] and scale space descriptor [23] approaches. Echoing the early hybrid SBVR systems of today, these early SBIR approaches relied on optimization processes that matched a sketch-derived model to each image — an expensive strategy that scales at best linearly with database size. Global image descriptors such as Edge Histogram Descriptor or Tensor Structure [13] allowed for improved retrieval times using efficient indexing, rather than model fitting matching strategies. Following the success of feature space quantization and the *Bag of Visual Words* (BoVW) paradigm for large scale image search, SBIR became scalable to large scale databases through the work of Hu et al. [15] and Eitz et al. [14] who proposed new descriptors adapting the BoVW for SBIR. Cao et al. later addressed SBIR scalability through inverse (assocative) index structures such as the Edgel Index. Sun et al. [30] achieved indexing of billions of images by optimizing this indexing structure using feature quantization and hashing strategies. In contrast to SBIR, SBVR has been explored sparsely. Initial systems such as Chang et al's VideoQ [8] echoed early SBIR work through the representation of objects as colored blobs, additionally incorporating indicators of object motion within the sketch. Collomosse et al. [9] over-segmented video super-pixels via a Linear Dynamical System (LDS) framework overcoming the requirement of an 'ideal video segmentation' assumed by VideoQ's video ingestion pre-process. Hu et al. [16] matched clustered SIFT correspondence tracklets although suffered in accuracy by disregarding shape and other appearance information such as color. We note that all these prior SBVR systems eschew the creation of efficient indexes for a model fitting approach in order to achieve usable accuracy.

The integration of semantics and appearance within SBIR (so called 'hybrid SBIR') has become an increasing trend, and has enabled SBIR to scale to billions of images. Approaches such as [5, 31], utilize semantics to find relevant images via tags providing representative imagery to remove ambiguity in the sketch matching process. Hybrid SBVR has also been explored by Hu et al. [19], extending their earlier trajectory retrieval system [16] to allow for semantic label annotation in the query sketch. Markov Random Fields (MRF) [17] were adapted for SBVR echoing the optimization approach of Collomosse et al. [9] for SBIR, but labelling space-time subvolumes rather than per-frame superpixels. This greatly improved the performance of retrieval, yet Hu et al. MRF approach still suffers from high computational expense requiring minutes to execute search queries over only a few hundred videos.

Relevance feedback (RF) is a popular technique for trying to bridge the *semantic gap*, improving retrieval results by inviting the user to identify relevant and/or irrelevant results that can then be used to refine subsequent retrieval iterations. The first RF techniques were presented for text retrieval by Salton et al. [25], and were adapted to Image retrieval by Su et al. [29]. Common approaches apply a linear SVM to rerank results [26] more recent approaches are applied over multiple features using multiple classifiers[32] or kernels [33]. Within SBR, image based RF techniques have been demonstrated [22, 12] yet, despite the well-known advantages of RF when dealing with complex multi-model datasets, RF has not yet been explored for SBVR.

3. SKETCH BASED VIDEO RETRIEVAL

Our system accepts a keyword annotated free-hand sketch $\mathcal{Q}$ as a query, and searches a database of video clips $\mathcal{D} = \{V_1, ..., V_D\}$ to find a match in real-time. Sketches are assumed to coarsely approximate object shape and color, the latter using the standard Macintosh 16 hue palette. Each sketched object is annotated with both a keyword describing its semantic category, and an arrow indicating motion trajectory, examples of this can be seen in Fig. 7. Shapes in the background of the scene may also be depicted using color and semantics.

3.1 Video Ingestion

Each video V_i ingested into the database is subject to a series of pre-processing steps. Each frame $V_i(t)$ is first over-segmented using mean-shift [11] into sets $\mathcal{S}_i(t) = \{S_1^t, ..., S_m^t\}$ of super-pixels; typically $m \simeq 200$ with minimum area $|S_i^t| = 20$ pixels. Histograms describing color $C(S_i^t) \in \Re^{16}$ and semantic attributes $Q(S_i^t) \in \Re^{12}$ are extracted for each super-pixel (Sub-sec. 3.1.1). A set of moving feature points $\mathcal{P} = \{P_1^t, ..., P_n^t\}$ corresponding to foreground objects are also identified within each frame, in a camera motion compensated space (Sub-sec. 3.1.2). These set of super-pixels and feature points are subsequently combined (with information from the background) to form our spatio-temporal descriptor for matching (Sub-sec. 3.2).

[1] Dataset available at http://cvssp.org/projects/sketch/cvmp14/

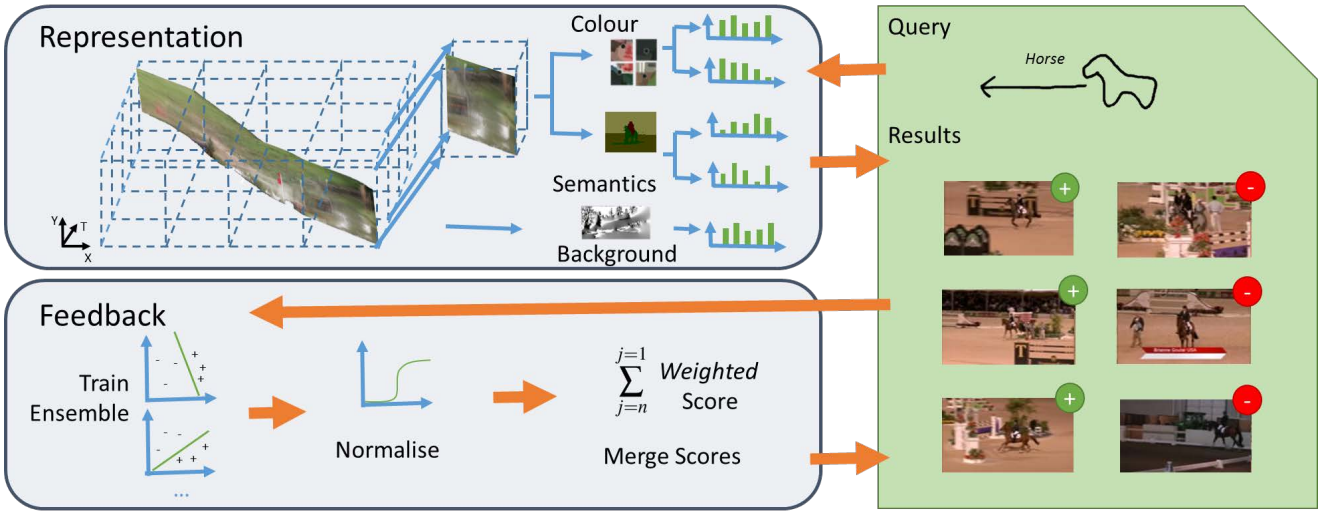

Figure 1: Videos and Queries are represented through space time volume by translating points into a panorama space. The volume is quantized into bins, with each bin having a semantic descriptor averaged over the total semantic distribution, and a color histogram from points within the cell. BoVW GF-HOG descriptor is generated over the panorama to represent the background structure. Results are provided back to the user for annotation of feedback. Linear classifiers are trained based on this feedback and combined to rerank the results.

3.1.1 Attribute Extraction

We extract semantic attributes from video by labeling pixels in each frame independently using Semantic Texton Forests (STF) [27]. Although more accurate and spatially coherent approaches exist [28, 21] their reliance upon complex filter banks and assignments at test time are prohibitive for scaling over a large video dataset. STF uses a Forest of Extremely Randomized Trees to classify pixels, these ensembles of decision trees are fast to train and test whilst their inherent randomness allows for flexibility to inter-class discrepancies.

The STF approach is composed of three classifiers: a standard ensemble of randomized decision trees; a global image classifier; and a second ensemble based on region information. The standard ensemble of trees are trained based on CIELab color value differences. A random point p_{x_2,y_2,b_2} around a training point p_{x_1,y_1,b_1} within a window ($width = 50$) is selected, where x,y refer to location and b refers to color channel. The comparisons of values are based on a random comparison function value p_{x_1,y_1,b_1} addition $p_{x_1,y_1,b_1} + p_{x_2,y_2,b_2}$, subtraction $p_{x_1,y_1,b_1} - p_{x_2,y_2,b_2}$ and absolute difference $\left|p_{x_1,y_1,b_1} - p_{x_2,y_2,b_2}\right|$, as in [27]. We also utilize $median_{p \in x}$ $median_{p \in y}$ as well as relative xy positions to centre $\frac{p_{x_1,y_1} - c_{x,y}}{p_{x_2,y_2} - c_{x,y}}$ where $c_{x,y}$ is the coordinates of the center of the image. In experiments these additions are beneficial due to salient objects commonly being in the center of shot.

The global image classification is computed using an approach similar to Bag of Visual Words(BoVW). A hierarchal comparison of leaf distribution of the decision trees forms a Pyramid Matching Kernel for use within a OneVsOthers SVM strategy. The third classifier, a second ensemble of decision trees trained on the resultant soft classified image. The probability image is subsampled and integral images are calculated allowing for a superpixel representation, a second ensemble is trained using rectangle feature is used based on window summation.

Each individual pixel is thus attributed a 12-D vector of probabilities describing the semantic content it depicts; this vector is averaged for all pixels within each superpixel $\mathcal{S}_i^t$ to yield $Q(\mathcal{S}_i^t) \in \Re^{12}$. This is in contrast the more common approach of conditional random fields[28], we opt for this approach for performance.

We extract color attributes by calculating a color histogram from pixels within each super-pixel $\mathcal{S}_i^t$. A histogram representation requires quantization of the color space into bins, for which we use the 16 color palette of the query sketch interface to enable rapid comparison at query-time. As this palette differs from the dominant colors present within each video, a remapping is performed. Given a video we collate pixels from several frames sampled at equal temporal intervals, and quantize CIELab space to identify the set of dominant video colors $V = \{v_1, ..., v_q\}$. Given a super-pixel region $\mathcal{S}_i^t$, we produce a normalized histogram $H_v(j), j = [1,q]$ by counting pixels within the region and quantizing using the video palette. The color descriptor $C(\mathcal{S}_i^t) = H_q(i), i = [1,16]$ for the super-pixel (where H_q is defined using the query palette) is given by:

$$H_q(i) = \frac{1}{|V|} \sum_{j=1}^{|V|} H_v(j) d\left(h_c[i], h_v[j]\right). \quad (1)$$

where $d(.)$ is the normalized CIELab distance between colors corresponding to the i^{th} and j^{th} bins of H_q and H_v.

3.1.2 Moving Feature Detection

SBVR queries depicting motion often do so relative to the background [10]. We must therefore compensate for background motion and do so by computing inter-frame homographies (after [7, 9]) to register the location of $\mathcal{S}_i^t$ within a single static reference frame. The bounding boxed surrounding all $\mathcal{S}$ in this space is used to scale the frame (preserving aspect ratio) to a constant width. This scaled, bounded region is in turn mapped to the rectangular region of the query canvas by a further scaling in order to transform points in the sketched query to points in the camera-compensated video space.

We extract a set of 2D feature points from the centroids of the superpixels $\mathcal{S}$, mapping the location of these into the camera-compensated space. We assume only moving objects to be of interest, and create a mask of moving regions using a thresholded optical flow field [4] (computed in the camera-compensated space). Feature points falling within the mask are stored for each frame resulting in $\mathcal{P}_{\{\}} = \{P_1^t, ..., P_n^t\}$.

3.1.3 Capturing Background Detail

We compute a single set of semantic and color attributes for the background region, i.e. all super-pixels within the camera-compensated space that do not fall entirely under the flow-derived motion mask of Sub-sec. 3.1.2. Referring to the union of such super-pixels as $\mathcal{B}^t$ we average the semantic and color distribution of all member super-pixels to obtain $C(\mathcal{B}^t)$ and $Q(\mathcal{B}^t)$.

Although background information can be depicted through blobs contributing to the total distributions, this is a cumbersome method and doesn't define query topology. We opt for an additional representation, taking the shape descriptor of Hu et al. [18] over the video panorama. Utilizing such an approach allows for dominant lines, such as horizon or mountain to be depicted in a consistent style.

We build a Bag of Visual Words over the dataset. Canny edge points guide the construction of a poisson gradient field. The area around these points is then described through Multi-Scale Histogram of Gradients. Following Hu et al. findings we build a vocabulary of 1500 words, video clip descriptors are then hard assigned to the vocabulary. The frequency histogram is normalised according to TF-IDF and we write the resultant histogram as $A(\mathcal{B}^t)$. To avoid encoding the layout of the panorama, we remove points that fall within 10 pixels of the boundary. Additionally for computational efficiency panoramas are resized to a fixed with of 500 while maintaining the aspect ratio, this makes it possible to compute the background in an efficient time.

3.2 Spatio-temporal Video Descriptor

In order to index a video we compute a descriptor from its spatio-temporal (x,y,t) volume. The volume is subdivided equally into cells; we empirically find a coarse quantization level of $6 \times 6 \times 6$ divisions for each dimension respectively to yield the best results. Each video V results in five histograms representing independent facets, foreground semantics, background semantics, foreground color, background color and background structure. Foreground facet histograms are formulated using knowledge of $\mathcal{S}$, $\mathcal{P}$, $C(.)$, and $Q(.)$ from pre-processing. Feature vectors computed independently for each cell are concatenated together, and an additional feature vector representing the video background is also appended.

3.2.1 Cell descriptor

To compute the feature vector for a given cell, we first identify the subset of feature points $\mathcal{P} = \{P_i^t\}$ falling within its spatio-temporal bounds $p \subseteq \mathcal{P}$ and the associated super-pixels that these points belong to $s(p) \subseteq \mathcal{S}$. We then compute a normalized color histogram from those superpixels:

$$H_c = \frac{1}{|p|} \sum_p C(s(p)). \tag{2}$$

A distribution of semantic attributes present in p is similarly computed but not normalized; in the case of color we are interested in the relative color distribution over all points present, whereas with semantic attributes we are interested in the total evidence for each semantic category trained.

$$Q_p = \sum_p Q(s(p)). \tag{3}$$

Cell descriptors are then concatenated.

3.2.2 Background descriptor

The feature vector for the background is a concatenation of distributions $S(\mathcal{B}^t)$, $Q(\mathcal{B}^t)$, and $A(\mathcal{B}^t)$.

3.3 Construction of Query Descriptor

We employ a sketch parsing step similar to [10] to extract individual object shapes from the sketch; full details are outside the contribution of this paper. The method results in a set of regions corresponding to the background and each foreground object, with 2D trajectories across the canvas associated with the latter. Each segmented region has a color distribution and may also have semantic label associated with it by the user.

We construct a spatio-temporal descriptor from the query sketch, to enable direct comparison with the spatio-temporal descriptor of each video in the database, as follows.

We first synthesize a set of super-pixels $\mathcal{S}$ and feature points $\mathcal{P}$ from the sketched regions corresponding to foreground objects. We assume that a sketched object progresses linearly along its sketched trajectory, for the duration of the video, with the sketched position being the start position. This yields an idealized position for the object at any relative time in the video. When synthesizing the position of the object, we use the coordinate mapping established between the sketch canvas and (constant width) camera-compensated video space to determine the region occupied by the object at each frame.

On this basis we synthesize a spatio-temporal representation of an 'ideal' video clip, generating $\mathcal{S}, \mathcal{P}$ progressively at each time instant in the 'ideal' clip. We cannot know the duration of this ideal clip, however this does not matter and can be arbitrary as we subsequently compute a descriptor ($6 \times 6 \times 6$) spatio-temporal quantization over the ideal clip duration — extracting a spatio-temporal descriptor as per (Sec. 3.2). Background properties are extracted from colors, labels and shapes on the sketched background as per Sub-sec. 3.1.3

3.4 Matching

Given the common representation of the query and video spatio-temporal descriptors, matching can be achieved trivially via Euclidean distance for each video descriptor in the database. Independently computing distances between the semantic and color components (using the Euclidean and χ^2 distances respectively) yields a performance gain of $\sim 10\%$. To retain the efficiency of computing a single norm between query and video descriptors, we borrow Arandjelovic and Zisserman's [1] trick of square-rooting each bin value (here, the color histogram bins) to convert Euclidean distance within the color sub-space to the Hellinger distance.

Sub-spaces of the descriptor could be rescaled (re-weighted) to reflect user preference U_w for one modality (e.g. semantics) over another (e.g. color); however equal weighting has been used in all results presented here.

3.5 Relevance Feedback

The significant performance benefits of our index-based matching approach for SBVR enable near-instantaneous full database search over hundreds of videos (Sec. 4). This raises the opportunity of working with the user 'in the loop' to interactively refine results. After candidate results are presented via our initial matching process (Sec. 3.4), the user is invited to label results indicating a few positive (relevant) or negative (irrelevant) examples. Results are then re-ranked using this input; a process referred to as Relevance Feedback (RF). In classical information retrieval RF is implemented by training an SVM within the descriptor space, using the relevant and irrelevant results labelled by the user. However unlike classical contexts, our hybrid sketches exhibit multiple modalities (or 'facets'; namely color, shape, motion, semantics, and background structure). We therefore implemented RF via an ensemble of classifiers, one per facet.

Each facet's classifier is trained within the sub-space of our spatio-temporal descriptor relevant to the facet; e. g. the color components of our descriptor form a sub-space in which the classifier for the color facet is trained. Each facet's classifier takes the form of a linear SVM ($\mathcal{M}_i$) trained using the marked up relevant and irrelevant results. A confidence weight $\mathcal{C}_i$ is assigned to each facet's SVM, estimating the discriminative power of that facet for the current query. For each facet samples specified as positive or negative become the training set $\mathcal{X}$ with their respective labels $\mathcal{Y}, \in [-1,1]$. The trained SVM $\mathcal{M}_i$, then yields weight $\mathcal{C}_i$ as:

$$\mathcal{C}_i = \frac{1}{n}\sum_{j=n}^{j=1} \begin{cases} 1 \text{ if } sgn(N(\mathcal{M}_i(\mathcal{X}_j))) = \mathcal{Y}_j \\ 0 \; otherwise. \end{cases} \quad (4)$$

Where $\mathcal{N}(.)$ normalises the kernel score using a sigmoid function:

$$N(\mathcal{M}_i(\mathcal{X}_j)) = \frac{1}{1+\exp(Af+B)}. \quad (5)$$

where A and B are optimized by a sigmoid fitting function using the method of Platt [24].

The position of a video in the re-ranked results is simply the product of each facet's SVM decision function $N(\mathcal{M}_i)$ and the confidences in that facet being discriminatory (useful) for the query in hand, determined automatically $\mathcal{C}_i$ and via the user defined weighting U_{w_i} of Sec. 3.4.

On each iteration of relevance feedback, the user-supplied relevant and irrelevant examples augment the training set and $\mathcal{M}$ are re-trained. During evaluation we found that presentation of only 10-15 results to be necessary to achieve significant improvement within a couple of iterations of relevance feedback (Sec. 4.3).

4. RESULTS AND DISCUSSION

We evaluate our over a dataset of 700 TV broadcast sports video clips, extending the datasets used by [9, 17]. Clip duration is 4-10s at 25fps with a mixture of low resolution PAL (720x576) and HD (1920x1080) footage. A ground-truth is manually defined noting the direction, semantic class and color of moving objects present. In our experiments we use twelve semantic classes {*person, horse, car, grass, snow, road, sky, trees, stands, obstacle, water, sand*}.

Retrieval time using our system is 3 orders of magnitude lower than prior hybrid SBVR techniques taking on average 150ms to linearly search the entire dataset.

Method	Proposed	[9]	[16]	[17]
Speed Per Clip (s)	0.0003	0.24	0.02-0.03	0.10684
Dataset Speed (s)	0.15	120	17.5	74

In contrast Hu et al. [15] took 20-30ms per clip approximately 17 seconds for our dataset, Hu et al. [17] took well over 1 minute for our dataset. Additionally our technique and attributes of our descriptor scales sub-linearly via the use of a $kd-$tree (disabled here for fairness of comparison).

4.1 Query Matching

Accuracy of our system is illustrated in Fig. 4, containing Average Precision-Recall (P-R) curves for 12 query sketches encompassing 7 different object colors and 8 object trajectories. A mean average precision (MAP) of 35% is achieved using matching semantics, shape and motion cues alone – falling to 32% when color is also incorporated. A similar small drop in performance is reported in other hybrid SBVR systems [17] due to the increased difficulty in accurately matching across all query modalities.

Fig. 8 illustrates representative queries of foreground objects. We observe that motion alone queries are easier to match. The car class is most challenging, this is due to the difficulty in semantic segmentation. Since the approach used for semantic segmentation (STF See Sec. 3.1.1) is based color, sports cars of which are often made up of a mixture of colors with little consistency between clips so challenging this method of segmentation. However opting for this algorithm enables us to segment frames in just under 1 second per frame, as opposed to TextonBoost [28] taking 7 seconds; this on the scale of our dataset constitutes a saving of over one week of time on video ingestion and is in line with existing Hybrid SBVR approaches [16, 17].

For comparison we adapt our ground-truth to the more permissive methodology of Hu et al. [17] (matching only 4 major directions) and compare using their dataset using 7 queries similar to those in their paper. We achieve MAP of 30% versus their 48%. Although accuracy is lower, our approach is 3 orders of magnitude faster and can scale sub-linearly whereas [17] scales linearly and comprises an expensive matching function is already intractable for interactive retrieval taking over a minute for our dataset.

4.2 Background

Structural information is a novel facet not explored by previous Hybrid SBVR [19, 17]. We demonstrate the results of an exemplar query that takes a simple background structural element into account. To visualize where the structure has come from we visualise the video panorama and overlay a plausible matching. We show these results in Fig. 3.

4.3 Relevance Feedback

For relevance feedback we take the 12 queries as outlined in Sec 4.1 and pass them into our system as normal. Taking the top 15 results of each query, they are then judged for relevance according to the ground truth. The relevance of resultant videos is fed back into the system for relevance feedback as in Sec. 3.5; this constitutes one iteration. Further iterations are based on the updated results, optimizing the ranking to the user request. Fig. 5 demonstrates the significant performance benefits achievable with just one iteration of RF. A gradual improvement from 32% to 50% MAP is observed over 4 iterations of feedback. Generally a satisfactory result, com-

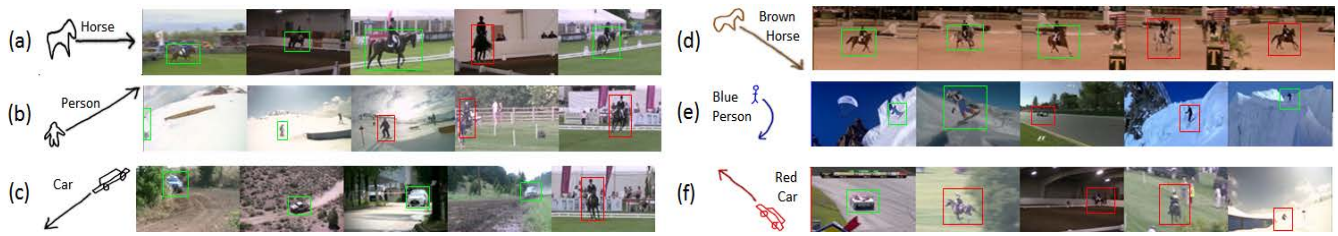

Figure 2: Five representative hybrid SBVR queries and top 5 results: (a)-(c) specify semantics and motion, (d)-(f) specify semantics, colour (shown as keyword for clarity) and motion. Shape is specified implicitly by sketch in both. Objects correctly identified are highlighted with green bounding box, red for incorrect results.

fortably exceeding the state of the art, can be obtained in just a couple of RF iterations.

	Original	Iter. 1	Iter. 2	Iter. 3	Iter. 4
All Modalities	32.1	42.3	45.5	49.3	49.7
Without Color	35.0	33.5	42.9	50.8	54.4

For clarity we demonstrate the improvement over the top 10 results for the 4 iterations over a couple of queries in Fig. 6b. We demonstrate over two of the more challenging queries, including a car query that was earlier highlighted from 8 as difficult, demonstrating that it is possible to overcome the challenges of a incorrect semantic segmentation during pre-processing, recovering 4 in 10 correct results from an original 1 in 10 in this difficult case.

	Original	Iter. 1	Iter. 2	Iter. 3	Iter. 4
Hu [17]	48	*	*	*	*
Proposed	30.6	38.0	46.4	**50.7**	**52.8**

As in Sec 4.1 for comparison we apply the ground truth of Hu et al. [17]. Applying our relevance feedback approach achieving an improvement of 5%, achieving 53% in contrast to 30%. For each iteration we achieve 30.6, 38.0, 46.4, 50.7, 52.8 for iterations 1,2,3,4 respectively. Demonstrating that 3 iterations are required to overcome the challenges of distilling a set of descriptors in contrast to model optimization, although a satisfactory top 10 result return can be achieved in fewer iterations.

An iteration of relevance feedback takes approximately 0.2 seconds, this includes the time to train and classify the datasets. This performance is comparable to the original query performance.

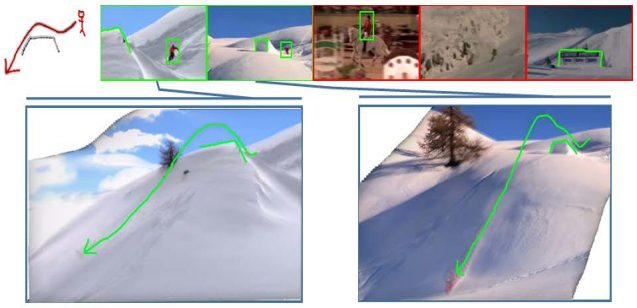

Figure 3: Top: Query incorporating background, and top 5 results; top 2 are relevant. Bottom: Video panorama depicting correctly retrieved object motion from best 2 results.

5. CONCLUSION

We have presented the first index-based *hybrid* SBVR system and utilized the interactive performance gains to introduce user guided relevance feedback (RF) to SBVR for the first time. Our system was driven by multi-modal free-hand sketches depicting various object facets such as appearance (color, shape), motion, background detail and the annotated keywords to indicate semantics. This is achieved using a set of novel spatio-temporal video descriptors. Prior hybrid SBVR builds a model from the sketched query, fitting this to each video via expensive optimization. Our initial indexing approach is 3 orders of magnitude faster using a linear search, with only minor loss of accuracy versus the state of the art optimization based approach [17]. However the scalability and performance (speed) gains of the indexing approach enabled RF through which we can improve upon the state-of-the-art accuracy by $\sim 5\%$. With each iteration of RF our results demonstrate we are able to greatly improve on the original results returned to the user.

Future work could consider variation on the components used, a particular challenge is the semantic segmentation process during video ingestion. However the use of RF to some degree mitigates error in these early decisions, and in the query itself, as the user can subsequently clarify their requirements. Due to the limited number of results we opted for a simple linear SVM for our ensemble RF approach. This might be extendable to an incrementally trainable SVM or a different approach for modelling the separation of relevant / not relevant feedback. We believe this machine learning question forms the most promising future direction for our work, though is not necessary to show the value of both our indexing approach and RF strategy to SBVR for video asset search.

6. REFERENCES

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[2] J. Ashley, M. Flickner, J. Hafner, D. Lee, W. Niblack, D. Petkovic, H. Sawhney, Q. Huang, B. Dom, M. Gorkani, D. Steel, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[3] A. Bimbo and P. Pala. Visual image retrieval by elastic deformation of object sketches. In *Proc. Intl. Symposium on Visual Languages*, volume 19, pages 121–132, 1997.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV 2004*. Springer Berlin Heidelberg, 2004.

[5] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. *CVPR*, June 2011.

[6] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: Interactive sketch-based image search on

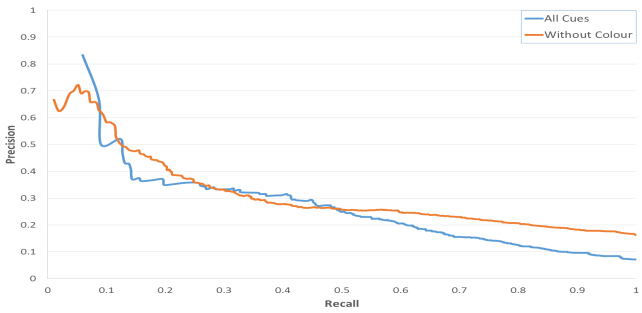

Figure 4: Average Precision-Recall curve for our system with all cues (blue) and without colour cue (red).

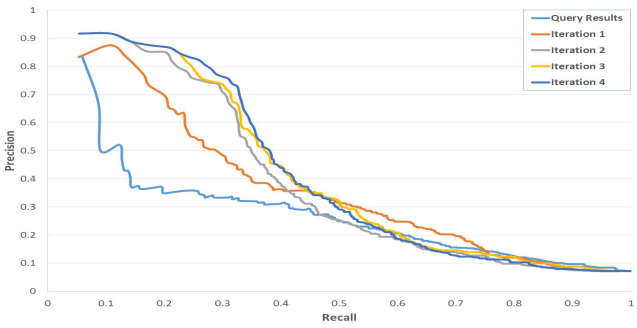

Figure 5: Average Precision-Recall curve over iterations of relevance feedback, the original query (light blue) results act as a baseline. Lines red,gray, yellow, dark blue refer to Iteration 1,2,3,4 respectively.

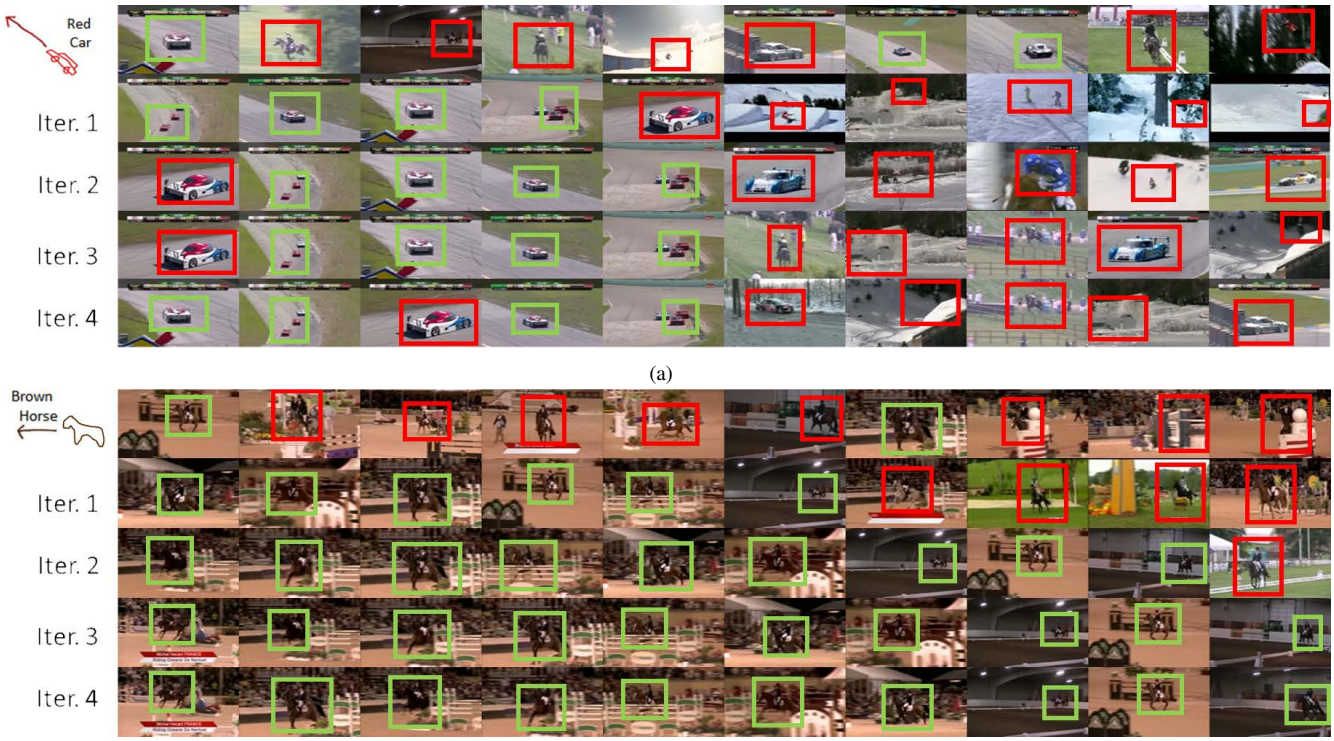

Figure 6: Top 10 results over four iterations for two different queries: (a) a red car identified as challenging in Fig. 2f, (b) brown horse

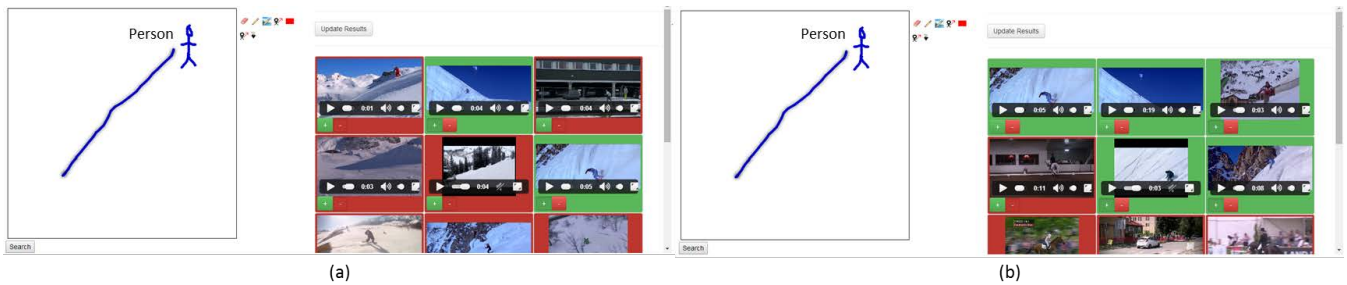

Figure 7: Original results of search based on blue stickman query on left (a), with updated results on right (b) after one iteration of relevance feedback. Videos are highlighted for relevance to query on both (a) and (b).

Figure 8: Collage of a subset of the dataset used, TV broadcast sports clips. Clips are of length 4-10s with a mixture of SD and HD resolution.

millions of images. In *Proc. ACM Multimedia*, pages 1605–1608, 2010.

[7] S. Chang, H. J. Meng, and D. Zhong. VideoQ: An automated content based video search system using visual cues. In *Proc. ACM Multimedia*, pages 313–324, 1997.

[8] S.-f. Chang, H. J. Meng, and D. Zhong. VideoQ : An Automated Content Based Video Search System Using Visual Cues. In *Proceedings of the fifth ACM international conference on Multimedia*, pages 313–324, 1997.

[9] J. Collomosse, G. McNeill, and Y. Qian. Storyboard sketches for content based video retrieval. In *Proc. of Intl. Conf. on Computer Vision (ICCV)*, 2009.

[10] J. P. Collomosse, G. McNeill, and L. Watts. Free-hand sketch grouping for video retrieval. In *Proc. ICPR*, 2008.

[11] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[12] E. Di Sciascio, G. Mingolla, and M. Mongiello. Content-based image retrieval over the web using query by sketch and relevance feedback. In D. Huijsmans and A. Smeulders, editors, *Visual Information and Information Systems*, volume 1614 of *Lecture Notes in Computer Science*, pages 123–130. Springer Berlin Heidelberg, 1999.

[13] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. A descriptor for large scale image retrieval based on sketched feature lines. *Proc. SBIM*, page 29, 2009.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Trans. on Visualization and Computer Graphics*, 17(11):1624–1636, 2012.

[15] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *Proceedings of Intl. Conf. on Image Proc. (ICIP)*, 2010.

[16] R. Hu and J. Collomosse. Motion-sketch based video retrieval using a trellis levenshtein distance. In *Proc. ICPR*, August 2010.

[17] R. Hu and J. Collomosse. Markov random fields for sketch based video retrieval. In *Proc. Intl. Conf. on Multimedia Retrieval (ICMR)*, 2013.

[18] R. Hu and J. Collomosse. A performance evaluation of the gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding (CVIU)*, 2013.

[19] R. Hu, S. James, and J. Collomosse. Annotated Free-hand Sketches for Video Retrieval using Object Semantics and Motion. In *Multimedia Modeling Conference*, page 12, 2012.

[20] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. In *SIGGRAPH*, pages 277–286, New York, New York, USA, 1995. ACM Press.

[21] L. Ladický and P. Kohli. Object-class segmentation using higher order CRFs. In *Proc. ECCV*, 2008.

[22] S. Liang and Z. Sun. Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recognition Letters*, 29(12):1733–1741, Sept. 2008.

[23] S. Matusiak, M. Daoudi, T. Blu, and O. Avaro. *Sketch-Based Images Database Retrieval*. 1998.

[24] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[25] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[26] L. Setia, J. Ick, H. Burkhardt, and A. I. Features. SVM-based Relevance Feedback in Image Retrieval using Invariant Feature Histograms. pages 542–545, 2005.

[27] J. Shotton, A. Blake, and R. Cipolla. Multiscale categorical object recognition using contour fragments. *IEEE Trans. PAMI*, 30(7):1270–81, 2008.

[28] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance , shape and context modeling for multi-class object recognition and segmentation. In *Proc. ECCV*, pages 1–14, 2006.

[29] Z. Su, S. Li, and H. Zhang. Extraction of feature subspaces for content-based retrieval using relevance feedback. *Proceedings of the ninth ACM international conference on Multimedia - MULTIMEDIA '01*, page 98, 2001.

[30] X. Sun, C. Wang, C. Xu, and L. Zhang. Indexing billions of images for sketch-based retrieval. *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 233–242, 2013.

[31] C. Wang. MindFinder : Image Search by Interactive Sketching and Tagging. In *Proceedings of World wide web*, pages 1309–1312, 2010.

[32] X.-Y. Wang, B.-B. Zhang, and H.-Y. Yang. Active svm-based relevance feedback using multiple classifiers ensemble and features reweighting. *Eng. Appl. Artif. Intell.*, 26(1):368–381, Jan. 2013.

[33] F. Yan, K. Mikolajczyk, and J. Kittler. Multiple Kernel Learning via Distance Metric. pages 147–156, 2011.