# A Sketch-Based Approach To Video Retrieval Using Qualitative Features

Koustav Ghosal Anoop Namboodiri
Centre for Visual Information Technology, IIIT Hyderabad, India
koustav.ghosal@research.iiit.ac.in
anoop@iiit.ac.in

## ABSTRACT

Motion trajectories extracted from certain videos contain sufficient spatio-temporal information which can be effectively used to characterize those videos. But the task of framing text-based queries for such videos in content-based video retrieval systems is very complicated. Sketch based query is an efficient tool to construct motion-based queries but perceptual differences like spatial and temporal variability pose serious challenges to query modelling.

In this work we propose a new method of modelling sketch based queries which attempts to extract the qualitative features of motion by minimizing the perceptual variability. We also develop a multilevel filter for indexing a query, in which the search results are refined at each stage using a cumulative scoring mechanism. Finally, we show the effectiveness of our algorithm on a dataset of real pool videos and a synthetic dataset containing simulated videos having very complex motion trajectories.

## Keywords

Content Based Video Retrieval, Sketch, Motion, Trajectories

## 1. INTRODUCTION

*Motion* has intrigued researchers in science and technology, sports, art, music, literature and films for ages. The trajectory of a missile, *Tiki-Taka* of Spanish football, the revolution of earth around the Sun, suspicious movements in railway stations — all these activities can be represented using a single or a combination of multiple motions. While motion itself conveys a lot of information for describing an event, depicting it (textually or pictorially) becomes a huge challenge for us. Depiction of motion in art has been there in five primary forms [1]- *Dynamic balance, multiple images, affine shear, blur, vectors.* Dynamic balance or broken symmetry deals with the pose of an object in an image from which the activity or event is predicted. A video can be summarized by overlaying key frames on one another and creating stroboscopic images [2]. (Figure 1 (b)). Affine shear and blur are two well known methods that are used to represent motion in graphic engines and comics. Vectors, on the other hand, are closest to human perception when it comes to representing motion [3].

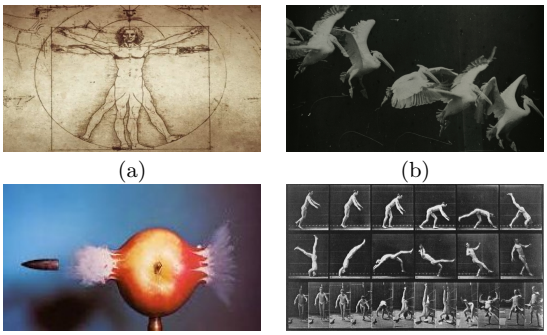


**Figure 1: Masterpieces of the past [2] (a) *Vitruvian Man* by Leonardo Da Vinci (b)*Flying Pelican* by Étienne Jules Marey (c) *Bullet* by Edgarton (d) *Headspring* by Muyibridge**

In Computer Vision, motion has been used as the primary *content* of a video in Content Based Video Retrieval Systems(CBVRS). The motion-based analysis of a video is frequent in surveillance, human-machine interaction, automatic target recognition and automotive applications [4]. Most existing approaches have primarily two phases in a CBVRS pipeline. In the first phase, trajectories of different objects are extracted from the videos and are stored. In the next phase, when a query (example videos, keywords or sketch) is presented to the system, it is matched with all the stored trajectories in the database and the corresponding video is retrieved. A serious drawback with example based queries is that an example is not always available in real time scenarios. Text based queries, on the other hand, are not suitable to describe long and complicated motions. For example, queries like "*the first strike in carrom where three or more carrom men or disks go to pockets*" or "*a particular diving style in swimming where the swimmer does three somersaults before diving*" are very difficult to frame. In addition to these problems, text based queries mainly look for texts associated with the video in the form of metadata or speech transcripts [5] instead of the visual content.

[2]Image Courtesy : www.google.com


*ICVGIP* '14, December 14-18, 2014, Bangalore, India

http://dx.doi.org/10.1145/2683483.2683537

A user-sketch can be used as an effective tool in such scenarios. But it involves a different set of challenges. The user perception (sketch) of a video is only an abstraction of the same. All the properties of a trajectory like shape, length and position are merely approximations of the trajectory of the object in the video. A simple Euclidean distance match is not bound to yield any meaningful result. Apart from spatio-temporal variability, the different sketches of the same trajectory by different users also suffer from perceptual variability. In other words, humans perceive motion in a way that is *qualitatively* similar but differ *quantitatively*. This is further elucidated by Figure 2.

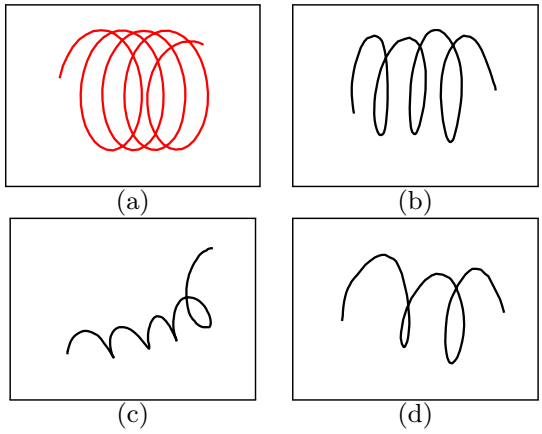



**Figure 2: (a) Original Motion Trajectory. (b) - (d) Interpretations of the same motion by different users**

So, the essence of the problem addressed in this paper lies in the question - Can we model the trajectories in a way such that the perceptual variability among different users for the same motion is reduced? In other words, can we define a space where the different instances of the user-sketch of the same trajectory are mapped similarly?

There has been a lot of work in trajectory extraction in the last decade, but as compared to that, modelling the user's perspective, according to our knowledge, is effectively not well explored. We have tried to study various aspects of this problem. Our work is organized as follows.

- In Section 3, a novel representation of a motion trajectory has been proposed which tries to remove the spatio-temporal variability among sketches of different users. Qualitative features have been derived, whose attributes tell us "how" rather than telling us "how much" about the different aspects of a motion.
- In Section 4, we propose an efficient multilevel cascaded retrieval method with a cumulative scoring mechanism, which boosts the retrieval accuracy at each stage of the cascade.
- In Section 5 we briefly describe our datasets and the intuition behind choosing them. We have conducted experiments on a real dataset of Pool videos and synthetic dataset of simulated videos.
- In Section 6, we present the experimental results using Precision-Recall Curves, Top-k accuracy and Mean Reciprocal Rank.

## 2. RELATED WORK

The existing research in this area can broadly be categorized into two different modules of a pipeline — *Trajectory Extraction* and *Query Indexing*.

In trajectory extraction, the objects are initially extracted from a key frame and then tracked across successive frames. Foreground segmentation in videos has been an extensively researched problem and several algorithms have been proposed for the cases of static [6], [7], [8] and dynamic backgrounds [9], [10], [11]. For tracking, standard techniques like Kalman Filters [12], Mean Shift Algorithm [13] and Double Exponential Smoothing [14] have been proposed. Once the trajectories are extracted, they are modelled by motion features like velocity, acceleration, curvature and length.

*VideoQ* [15], which is one of the first Content Based Video Retrieval Systems using sketch (sCBVRs), takes as an input a sketch, containing colour and shape based features and uses wavelet decomposition to model each trajectory. Alternative approaches to process trajectories like statistical modelling, Principal Component Analysis of sub-trajectories, MPEG based motion flow extraction methods have also been proposed. An exhaustive survey of these techniques can be found in Hu *et al.* [5]. This paradigm has been applied to the problem of event detection and activity classification as well [16]. Bashir *et al.*[17] and Cuntoor *et al.* [18] proposed HMM based approaches for trajectory based activity classification. Basharat *et al.* [19], Saleemi *et al.* [20] , Stauffer *et al.* [21] have modeled traffic behavior using spatio-temporal information from videos. Dyana *et al.* [22] have used a multispectro-temporal curvature scale space (MST-CSS) representation to describe a video object.

Inspite of the plethora of motion-trajectory based video retrieval systems, according to our knowledge, there are very few generic sCBVRs [15], [23], [24]. In [15], colour, shape and appearance of the objects have also been used for describing *content* of videos. These features work well when the database consists of videos that vary widely in content. On the contrary, if the videos are similar *e.g.* Pool or Billiards videos, ball trajectories have greater saliency than colour, shape *etc*, in terms of representing the video. Unlike surveillance videos, these trajectories are unconstrained with respect to direction and position. Sub-trajectory based matching as done by Chang *et al.* [15] is ideal for event search where the trajectories are short and number of sub-trajectories is limited. In case of longer trajectories, the temporal information is also important alongside spatial information and cannot be ignored.

Our work is inspired and closely related to the work by Bashir *et al.* [25]. They have represented trajectories as a temporal ordering of the sub-trajectories by using Principal Component Analysis, Spectral Clustering and String Matching. Like theirs, our work also relies on a stable trajectory extraction algorithm. But there are two fundamental differences between the two. Firstly, they have used *query by example* where the intention was to retrieve a similar set of trajectories from the database. Our query is *sketch based*, where the user intends to find an exact match. Since it is sketch-based, different users interpret the same motion according to their own perception, which differs quantitatively. Secondly, we have introduced a novel scoring mechanism that combines motion features like shape and direction in an efficient manner to refine the results.

## 3. MOTION FEATURES

We first define a feature representation *that captures the constraints among dimensions rather than their quantitative values* [26]. In Sections 3.1 and 3.2, we explain our strategy to model the sub-trajectories in user sketch and the original videos respectively. At the end of Section 3.2, we show how these sub-trajectories can be used to model the entire trajectory. In Section 3.3, we derive another set of features that represent directional characteristics of motion.

### 3.1 User Sketch

Query or sketch is obtained as a collection of $(x, y, t)$ points where $x$, $y$ and $t$ represent $x$ coordinate, $y$ coordinate and time respectively. While collecting data, the users were shown some videos, randomly sampled from the dataset and then asked to recollect as many motion trajectories from the videos as they can. But the match was carried out with the longest one among all the trajectories in a given video. In our case, longer trajectories were assumed to be more salient than shorter ones. In a different scenario, there could be other metrics to measure saliency.

The trajectory is first de-noised, freed from outliers and smoothened using the conventional spline interpolation [27]. A video may contain multiple motions. The trajectories are then normalized (empirical results showed that height, normalized to 100 gave best results) in such a way that the relative position and size of the trajectories remain intact. The aspect ratio of each trajectory is preserved. This relative normalization strategy gives the sketch translational invariance and also preserves their relative attributes. The trajectories are subsequently segmented based on curvature. We call each segment a *motion segment* ($m$-segment), a term inspired from *ballistic segment*, frequently used in modelling handwriting [28] (see Figure 3). Here, we assume that a motion can be random and unconstrained but the sub-trajectories follow a strict pattern. Our assumption is based on some fundamental principles of *rigid body mechanics* and *handwriting*. Unless interrupted by some external force, the $m$-segments are either linear or circular or parabolic in shape. If we consider each $m$-segment as an arc of a circle, then the corresponding *centre* and *radius* can be used to represent the arcs.

Each $m$-segment has the form: $S = \{[x_i, y_i] \mid i = 1, 2, \ldots n\}$. A circle is fit by minimizing the squared radial deviations, expressed as

$$J = \min_{x_0, y_0, r} \sum_{i}^{n} x_i^2 + y_i^2 - 2x_0x_i - 2y_0y_i + x_0^2 + y_0^2 + r^2 \tag{1}$$

where $[x_0, y_0]$ and $r$ is the centre and radius of the circle, respectively.

Let, $-2x_0 = a_1$,$-2y_0 = a_2$ and $x_0^2 + y_0^2 + r^2 = a_3$. Then Equation 1 can be expressed in matrix form as

$$(X\ Y\ 1)(a_1\ a_2\ a_3)^T = -(X * X\ +\ Y * Y) \tag{2}$$

where $*$ is the Hadamard product of two matrices and $X^T = [x_1\ x_2\ \ldots\ x_n]$ and $Y^T = [y_1\ y_2\ \ldots\ y_n]$ and $[x_i, y_i] \in S$. Solving Equation 2, we get

$$(a_1\ a_2\ a_3) = -(X\ Y\ 1)^+ \times (X * X\ +\ Y * Y) \tag{3}$$

where $P^+$ denotes the *Moore Penrose Pseudo-Inverse* of matrix P. Thus from the solution of equation 2, we can find our

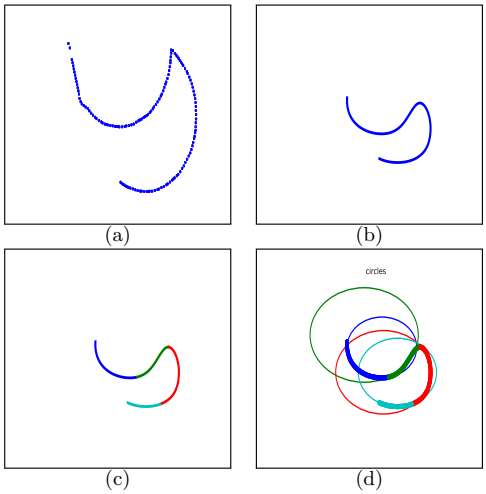



**Figure 3: A sample motion with the corresponding $m$-segments: (a) Original (b) Smooth and Normalized (c) $m$-segments (d) Circle-Based Representation**

desired circle parameters as

$$\begin{aligned} x_0 &= -\frac{a_1}{2} \\ y_0 &= -\frac{a_2}{2} \\ r &= \sqrt{\frac{a_1^2 + a_2^2}{4} - a_3} \end{aligned} \tag{4}$$

It is interesting to note that small, medium and large values for radius indicate circular, parabolic and linear motion respectively and giving us a qualitative understanding of the $m$-segment. The radius is mapped to $[0, 1]$ using a hyperbolic tan function. The notion of approximate position of segment $S$ can be represented using the mean $[x_\mu, y_\mu]$, which was experimentally found to be less variant than the centre of the circle. Subsequently, $S$ is represented as,

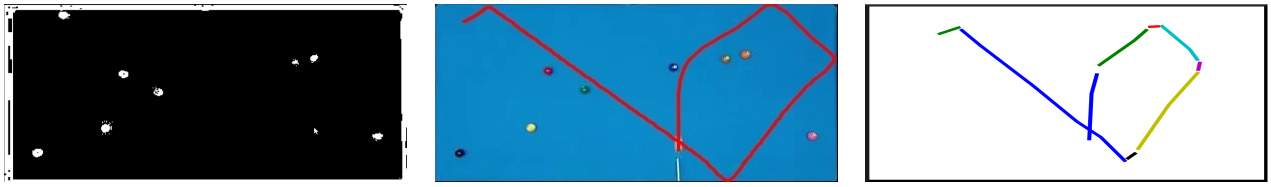
$$S = (\ x_\mu,\ y_\mu,\ r,\ m,\ s\ )$$

where $m$ is the slope of the best line fitting to the points in each segment. The parameters are estimated using a least squares approximation. The slope has been experimentally quantized to 8 directions (N,S,E,W,NW,NE,SW,SE) to minimize the perceptual variability. $s$ represents the normalized length of the arc.

### 3.2 Original Trajectory

In this work, our focus has been more on modelling user perception rather than trajectory extraction from videos. So we have used a set of 100 artificially simulated simple videos (Section 5) where the background is static and there are only a few objects. But the motion paths have been made very complex. We have also collected a set of 100 Pool Shot videos from many international matches, uploaded on YouTube. Motion trajectories were extracted from the real dataset in the following manner (Figure 4).

Firstly, we have denoised each frame using a median filter. Then we have extracted only the board region from each frame using a mask selected from the average frame. Next,



(a) (b) (c)

**Figure 4: (a) A Background Extracted Frame (b) Trajectory Extracted from a Video (c) Trajectory after smoothing and segmentation**

we have done background extraction using a thresholding based method. The moving components were tracked in the video using a Gaussian Mixture Model [8] over the binarized frames. The trajectories were extracted from the video using a Kalman Filter [12]. Multiple object tracking was implemented using a variant of Hungarian Algorithm [29]. The raw trajectories were pre-processed and the qualitative features for a segment *i.e* $S =< x_\mu,\ y_\mu,\ r,\ m,\ s >$ are obtained in a similar manner, as discussed in the previous sections. The $m$-segments extracted from all the trajectories in the database are clustered using the $k-means$ algorithm to obtain a codebook containing $k$ cluster centers.

The complete trajectory is modelled as a histogram of $m$-segments, with each bin of the histogram corresponding to each cluster center in the codebook. So, for each trajectory in the database we create a *bag-of-motions* representation, similar to the *bag-of-visual words* representation used to represent images with SIFT features [30]. This same codebook is used to generate the bag-of-motions representation for the query as well.

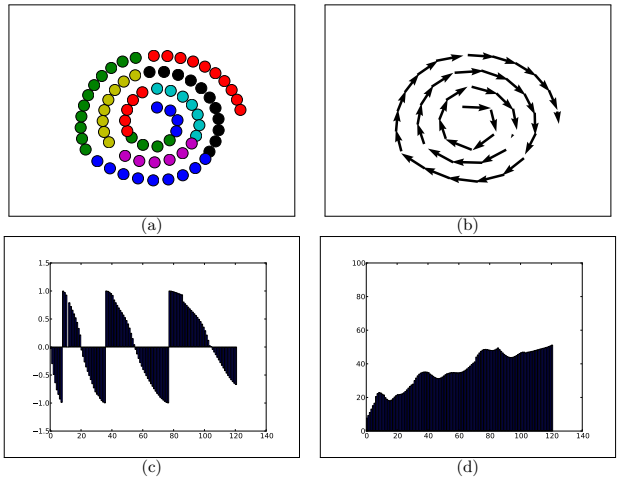
## 3.3 Order, Direction and Scale

Histogram based features proposed until now do not capture the important motion properties: *Temporal order, Direction and Scale*. As mentioned in Section 2, the order in which the sub-trajectories appear, plays a very important role in representing the motion. But it is difficult to compare two motion trajectories having unequal length. *Dynamic Time Warping* (DTW) [31] is an efficient tool which is used to compare time-series data having unequal length. We use DTW to compare the motion trajectories.

We find the change of direction across time. First we resample the trajectory to remove the variability in the density of points due to varying speed of the hand movements (Figure 3.3). Next, we divide a trajectory into equipoint segments (segments having equal number of points), and the distribution of direction across time has been approximated by fitting a line to each of the equipoint segments. The angle made by each equipoint segment with the horizontal $X$-axis has been mapped to $[-1, 1]$ using a *sine* function. (Figure 5(c)).

Similarly, scale of motion is an important factor for identifying motion trajectories. The scale is defined as the change of the current position with respect to the starting point. For example, a counter-clockwise spiral motion can be converging or diverging. Shape histogram or change of direction across time cannot differentiate between such motions. But the change of scale across time distinguishes the two.

Summarizing the previous discussions, it can be said that in our query we have conveyed four fold information. Firstly, we have conveyed shape information and the approximate



(a) (b) (c) (d)

**Figure 5: A Spiral Motion from our synthetic dataset (a) Points sampled equidistantly in each segment (b) Directions tracked for each equipoint segment (c) Temporal Change of Direction (d) Temporal Change of scale**

position of each segment in the trajectory in the *bag-of-motions* representation. Secondly, we have features that represent the change of direction and scale of the trajectory across time. Notably, none of the features we derive here use any absolute information. All the features approximate the overall motion, which is intended, to remove the perceptual variability among different users. These features give us a *qualitative* understanding of the motion trajectory.

# 4. RETRIEVAL

In this section, we propose our multilevel search strategy. Once a query is given, four sets of features described in the previous sections are extracted and the query is passed through a cascaded filter having four stages.

In the first stage the query histogram is matched with the database of *bag-of-motions*. Each sample $X_i$ in the database is assigned a score $\alpha_i^{(1)}$. This level of filtering finds the trajectories that have similar sub-trajectories. At the next level, a Dynamic Time Warping (DTW) [31] match is performed between the query $Q = [S_1, S_2, \ldots S_M]$ and each sample in database $T_i = [S_1, S_2, \ldots S_N]$, where each $S_i$ is the feature derived in Section 3.1 and 3.2. A new score $\alpha_i^{(2)}$ is obtained at this level. Apart from preserving the order, the match at this level also facilitates partial trajectory match. At the next two levels DTW match is carried out between the features derived in Section 3.3 and scores

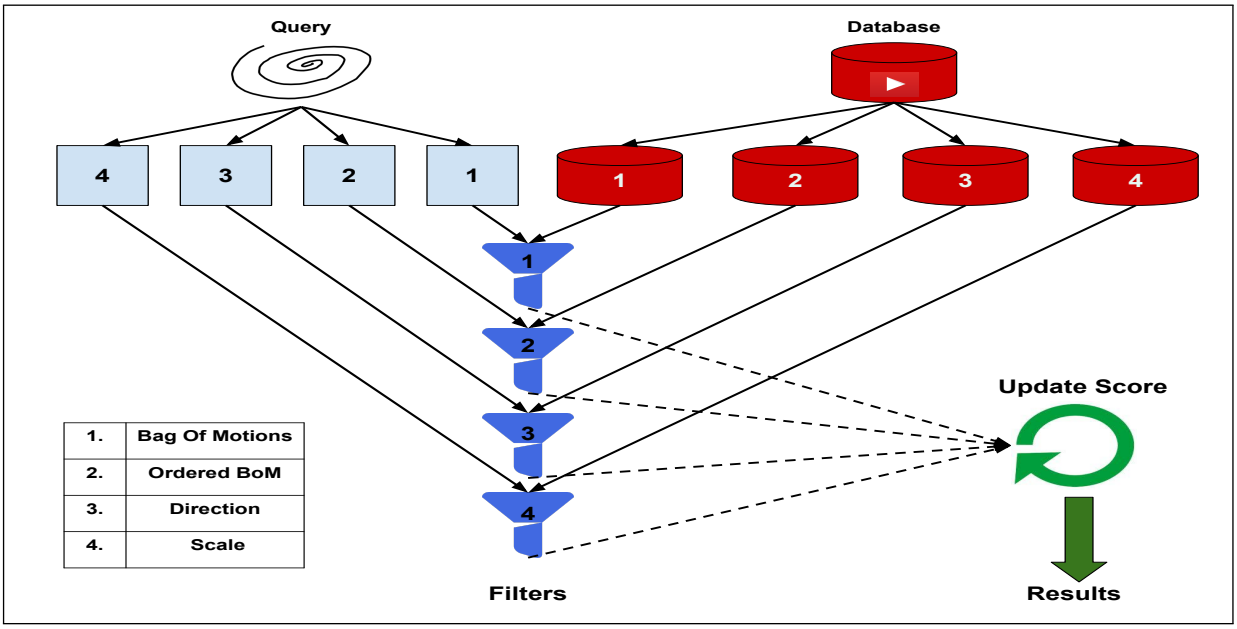



**Figure 6: Multilevel Retrieval Strategy : The query and original videos in the database ( top-left and top-right ) are processed and four sets of features are derived in each case. There are four different levels of filtering ( four blocks vertically arranged at the center ). The functionality of each filter has been shown in the table ( bottom-left ). After each level of filtering, the score is updated by the score update module ( bottom-right ). The videos are retrieved based on the final score.**

$\alpha_i^{(3)}$ and $\alpha_i^{(4)}$ are obtained. The final score $\alpha = \sum \alpha^{(i)}$ is calculated. Each of the scores are calculated as a function of the distance of the query from each sample, computed at each level *i.e* $\alpha^{(i)} = f(d^{(i)})$. The value of $\alpha$ is updated after every stage in a cumulative fashion. The final results are retrieved based on the value of $\alpha$ after the fourth stage. The algorithm has been explained using a block diagram in Figure 6.

Two important aspects should be considered here. Firstly, although multiple trajectories were extracted from a video, our current system uses only one trajectory sketch to search for the video. Chance of choosing a particular trajectory depends entirely on the user. But in our database we store all the trajectories. However, extending the system to allow a user to specify multiple trajectories in a video can further refine the results. We have not implemented this in our query interface as of now, but we intend to do this in our future work.

Secondly, our current system does not behave like a *regular* cascade and it differs from traditional cascaded systems. Currently, the search-space is not reduced at each level in our algorithm as it happens in cascaded detectors such as [32]. But please note that with successive stages, our features and matching go from weak and efficient to discerning and complex. We can discard samples with the lowest matching scores at each stage, making it a regular cascade.

## 5. DATASET

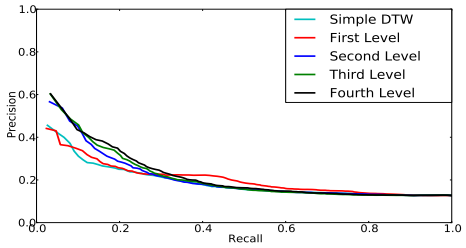
We have synthesized a dataset, which contains 100 videos of one, two and three body motions. The videos are divided into five sets : (a) a set with linear motions resembling Pool shots (b) a set with mixture of linear and exponential curves as trajectories that resemble moving cars and (c) a third set with respective motions like circular (clockwise and counter clockwise), sinusoidal and spiral. (d) a set of motions that resemble typical motions like sea-saw ride, people jumping side by side, divers diving *etc.* (e) where the motion trajectories are regular geometric shapes like square and triangle. It was found that in animation videos, most of the motion trajectories have regular geometric shapes. The synthetic dataset was created keeping in mind all the different kinds of videos which later can be explored with this kind of retrieval strategy.

We also tested our method on real pool videos. Full match pool videos were segmented into shots using a histogram based approach [33]. Then a dataset of 100 clean videos having a top view of the pool board was created. Each video was shown to different users and they were asked to group the videos which they found perceptually similar.

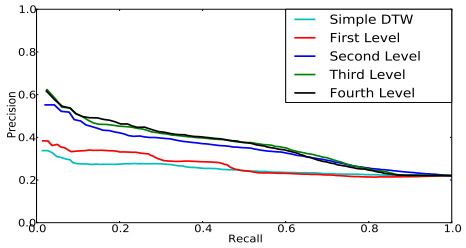
It was difficult to divide the pool shots into a specific number of classes. To achieve this, we asked multiple users to cluster/group the videos based on their similarity. Pairs of videos within a group were assigned a high similarity score and in different group were given a lower similarity. The similarity scores from multiple users were integrated into a single score matrix and an automatic clustering was performed to arrive at the final class divisions. It was found that the users could identify five different groups from the dataset. The distinction between different classes were done mainly based on shape, direction and position of the shots. The dataset is available for download and can be found online on our website[3].

For collecting query sketches, we sampled 50 videos from each of the datasets and then showed 20 videos (10 from each set) to a user. The user was asked to watch the videos care-

[3]http://cvit.iiit.ac.in/projects/sketchbasedretrieval/

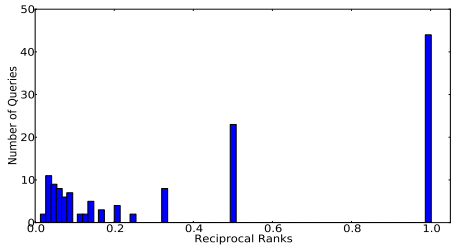



(a) Pool Videos dataset

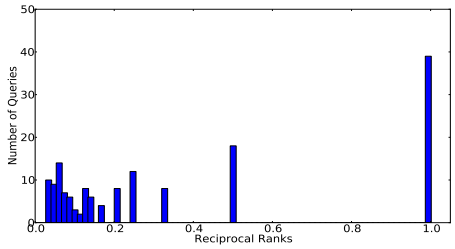



(b) Synthetic Motion dataset

**Figure 7: Precision Recall Curves (view in colour)**




(a)Pool Videos dataset




(b) Synthetic Motion dataset

**Figure 8: Reciprocal Ranks : A high value near one indicates that most of the queries retrieved the exact match as the first result. A value of 0.5 indicates that the second result was the correct match and so on.**

fully and then sketch two most salient motions that he/she could remember from each video. The $(x, y, t)$ coordinates of the sketches were recorded. The experiment was carried with 25 users. Each video had 5 sample queries.

## 6. EXPERIMENTS AND RESULTS

We have evaluated the effectiveness of our representation using three different standard evaluation metrics as follows.

**Precision Recall :** The PR curves generated from our experiments with real and synthetic datasets are shown in Figures 7 (a) and 7 (b) respectively. It can be seen that the area under the curve gradually increases as the query score gets updated with each filter. Simple DTW of the points performs worst. Only Bag-of-Motions based nearest neighbour search performs poorly (red curve). But the results improve significantly as soon as the temporal information is also used and the scores are updated in the next filter (blue curve). The precision is further tuned using the next filters and the best curve is obtained after the final stage of filtering is completed. There is a significant improvement after the second level than in third and fourth levels. This is because the order in which the sub-trajectories appear play a vital role in distinguishing motion trajectories. We believe the improvement reflects the importance of temporal ordering in modelling long trajectories. Also precision-recall curves in case of the synthetic dataset are better than those in case of the real dataset. This is mainly because, the synthetic dataset has more inter-class variance. The motions have fundamental differences with respect to shape and spatio-temporal properties. But in case of Pool Videos, the trajectories are mostly linear (except the trick shots) and have very little inter-class variance.

**Mean Reciprocal Rank :** We mentioned in Section 2 that our retrieval strategy is intended to find the exact match instead of a class of matches. We found Mean Reciprocal Rank as a good measure to test such an algorithm. The multiplicative inverse of the rank of the first correct answer in a set of retrieval results is obtained. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

We calculated the MRR with the same set of queries. It was found to be 0.5 and 0.4 on the real and synthetic videos respectively. It indicates that with this approach, the chance of finding an exact match within the top few results is high, which is desirable in our case. Figure 8 demonstrates the histogram of reciprocal ranks of all the queries.

**Accuracy :** We defined accuracy in the following manner. A search was considered successful, if the exact match appeared in the top $k$ results. Figure 9 demonstrates the accuracy values on the real and synthetic dataset for $k$ values ranging from 1 to 15. Figure 10 shows an example query with the top k retrieved videos.

## 7. CONCLUSIONS AND FUTURE WORK

In this work, we have addressed a lesser explored aspect of an extensively explored problem of motion trajectory based video retrieval.

One of the limitations of our algorithm is that it relies on strong foreground segmentation and trajectory extraction algorithms which are themselves unsolved problems in complicated videos. Challenges like dynamic background, camera shakes, shadow, camouflage *etc.* [34, 35] are active areas of research in Computer Vision. Moreover, this method cannot be used for retrieving videos where motion is not the most salient feature. The problem can also become very ill posed and difficult when the query is so complicated that it cannot be described in any unimodal format.

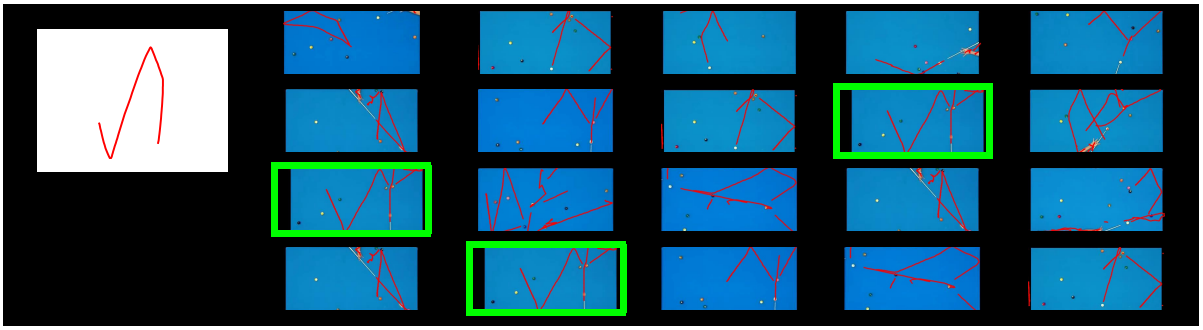
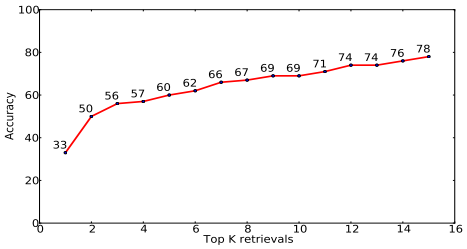
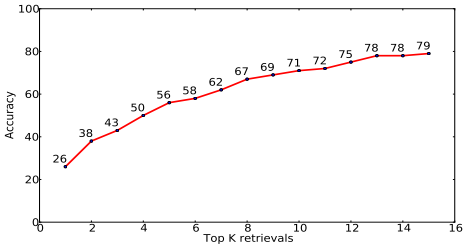


**Figure 10: The figure on the left is the query. On the right, the four rows correspond to the four stages of our filter. Elements in each row correspond to the top 5 results at each iteration, after the score is updated. The exact match is highlighted in green. At the first level, the exact match is not found in top 5. But it appears after Stage 2 and maintains its position within the top 5 results till stage 4**




(a) Pool Videos dataset




(b) Synthetic Motion dataset

**Figure 9: Accuracy at different top K retrievals. Exact accuracy values at different k are annotated on the curve. It can be observed that the accuracy reaches 70% within top-10 results for both the real and synthetic dataset.**

However, it is different from most of the existing sketch based systems because the query is unconstrained. No initial frame is supplied to the user. We have proposed a new representation for the trajectories of objects in videos and the sketch-based query. The features, which depend on perceptual similarity, are qualitative in nature and robust to user-level variations. Moreover, instead of using only spatial or temporal features the technique uses a combination of those by implementing a novel cumulative scoring mechanism.

A better understanding of the motion perception in humans will enable us to develop more robust features. But, our method can be applied on top of any trajectory estimation method. Also, it can be refined further with object level features like shape, colour and size. Also the fact, that we have shown that our method can be used on real Pool videos with satisfactory results, gives us hope that this approach can be extended to more complicated videos and used to develop accurate and robust multi-modal systems in future.

## 8. ACKNOWLEDGMENTS

We would like to thank all those from **Centre for Visual Information Technology, IIIT-Hyderabad**, who spent their valuable time to read and review this work since its inception.We are also grateful to our reviewers in ICVGIP 2014, who gave us thorough and insightful reviews for this work.

## 9. REFERENCES

[1] James E Cutting. Representing motion in a static image: constraints and parallels in art, science, and popular culture. *PERCEPTION-LONDON-*, 31(10):1165–1194, 2002.

[2] Carlos D Correa and Kwan-Liu Ma. Dynamic video narratives. *ACM Transactions on Graphics (TOG)*, 29(4):88, 2010.

[3] Wilson S Geisler. Motion streaks provide a spatial code for motion direction. *Nature*, 400(6739):65–69, 1999.

[4] Mattia Broilo, Nicola Piotto, Giulia Boato, Nicola Conci, and Francesco GB De Natale. Object trajectory analysis in video indexing and retrieval applications. In *Video Search and Mining*, pages 3–32. Springer, 2010.

[5] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.

[6] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31 Vol.2, Aug 2004.

[7] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[8] Andrew B Godbehere, Akihiro Matsukawa, and Ken Goldberg. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American Control Conference (ACC), 2012*, pages 4305–4312. IEEE, 2012.

[9] Antoine Monnet, Anurag Mittal, Nikos Paragios, and Visvanathan Ramesh. Background modeling and subtraction of dynamic scenes. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1305–1312. IEEE, 2003.

[10] V. Mahadevan and N. Vasconcelos. Background subtraction in highly dynamic scenes. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6, June 2008.

[11] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1778–1792, Nov 2005.

[12] Gary Bishop and Greg Welch. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8:27599–3175, 2001.

[13] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[14] Weiming Hu, Dan Xie, Zhouyu Fu, Wenrong Zeng, and Steve Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007.

[15] Shih-Fu Chang, William Chen, Horace J Meng, Hari Sundaram, and Di 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. ACM, 1997.

[16] Rui Hu, Stuart James, Tinghuai Wang, and John Collomosse. Markov random fields for sketch based video retrieval. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 279–286. ACM, 2013.

[17] Faisal I Bashir, Ashfaq A Khokhar, and Dan Schonfeld. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, 2007.

[18] Naresh P Cuntoor, B Yegnanarayana, and Rama Chellappa. Activity modeling using event probability sequences. *Image Processing, IEEE Transactions on*, 17(4):594–607, 2008.

[19] Arslan Basharat, Alexei Gritai, and Mubarak Shah. Learning object motion patterns for anomaly detection and improved object detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[20] Imran Saleemi, Khurram Shafique, and Mubarak Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8):1472–1485, 2009.

[21] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757, 2000.

[22] A Dyana and Sukhendu Das. Mst-css (multi-spectro-temporal curvature scale space), a novel spatio-temporal representation for content-based video retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(8):1080–1094, 2010.

[23] Jun-Wei Hsieh, Shang-Li Yu, and Yung-Sheng Chen. Motion-based video retrieval by trajectory matching. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(3):396–409, March 2006.

[24] Chikashi Yajima, Yoshihiro Nakanishi, and Katsumi Tanaka. Querying video data by spatio-temporal relationships of moving object traces. In *Visual and Multimedia Information Management*, pages 357–371. Springer, 2002.

[25] Faisal I Bashir, Ashfaq A Khokhar, and Dan Schonfeld. Real-time motion trajectory-based indexing and retrieval of video sequences. *Multimedia, IEEE Transactions on*, 9(1):58–65, 2007.

[26] Thomas F. Stahovich, Randall Davis, and Howard Shrobe. Qualitative rigid-body mechanics. *Artificial Intelligence*, 119(1âĂŞ2):19 – 60, 2000.

[27] G. Medioni and Yoshio Yasumoto. Corner detection and curve representation using cubic b-splines. In *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, volume 3, pages 764–769, Apr 1986.

[28] S.P. Teja and A.M. Namboodiri. A ballistic stroke representation of online handwriting for recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 857–861, Aug 2013.

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[30] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[31] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[32] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[33] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1(1):10–28, 1993.

[34] Thierry Bouwmans, Fida El Baf, and Bertrand Vachon. Background modeling using mixture of gaussians for foreground detection-a survey. *Recent Patents on Computer Science*, 1(3):219–237, 2008.

[35] Sebastian Brutzer, Benjamin Hoferlin, and Gunther Heidemann. Evaluation of background subtraction techniques for video surveillance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1937–1944. IEEE, 2011.