Towards SKETCHED Visual Narratives

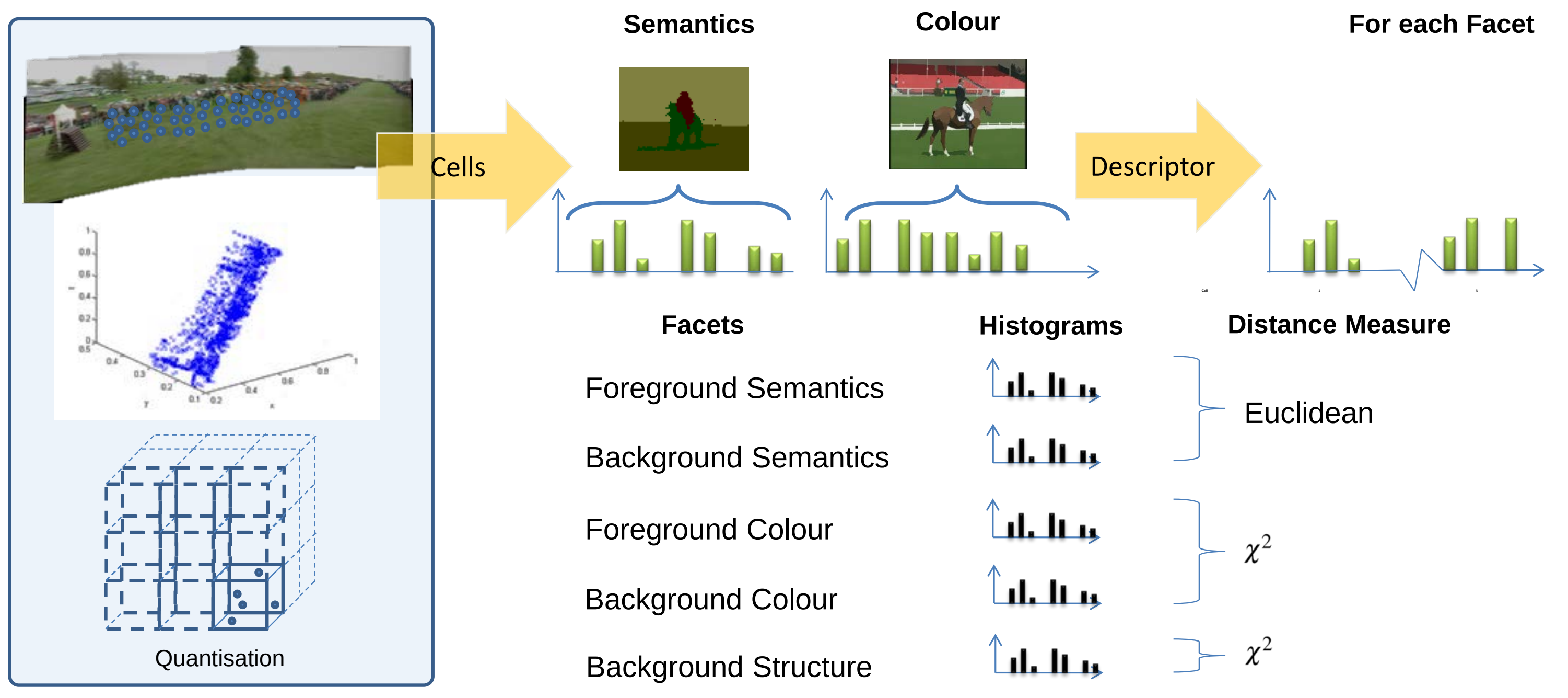
Stuart James, John Collomosse
Centre for Vision Speech and Signal Processing, University of Surrey
stuart@stuart-james.net

Introduction

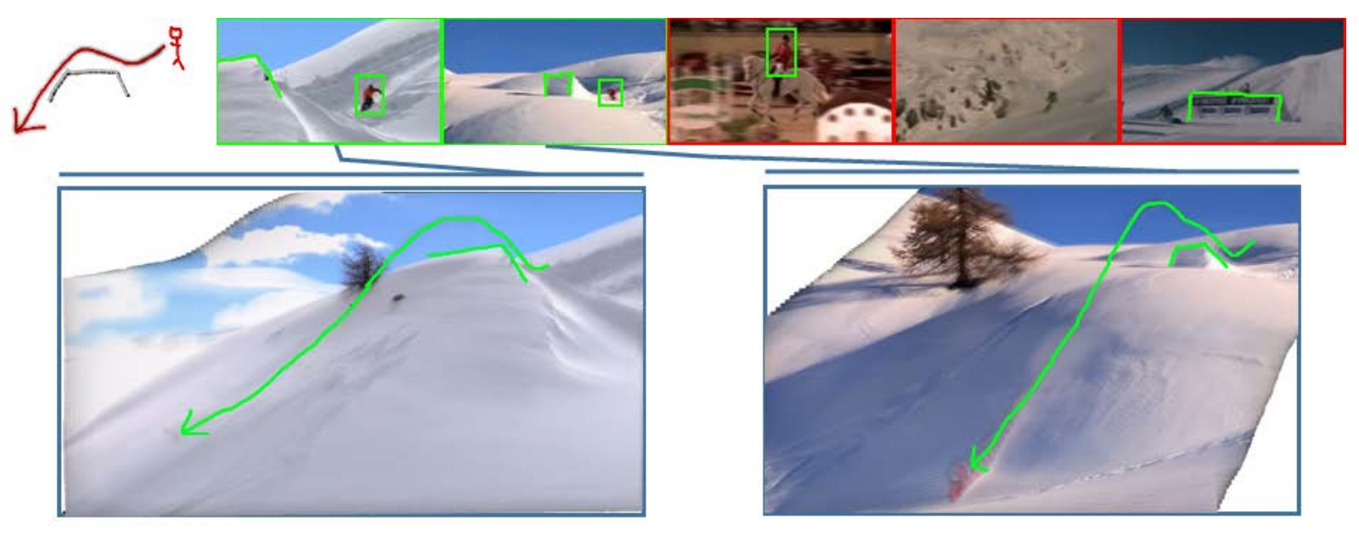
Humans have an innate ability to communicate visually; the earliest forms of communication were cave drawings, and children can communicate visual descriptions of scenes through drawings well before they can write. Drawings and sketches offer an intuitive and efficient means for communicating visual concepts. We present several new algorithms for searching and manipulating video using free-hand sketches. We propose the Visual Narrative (VN); a storyboarded sequence of one or more actions in the form of sketch that collectively describe an event. We show that VNs can be used to both efficiently search video repositories, and to synthesise video clips.

Efficient Annotated Sketch-based Video Retrieval [1,2]

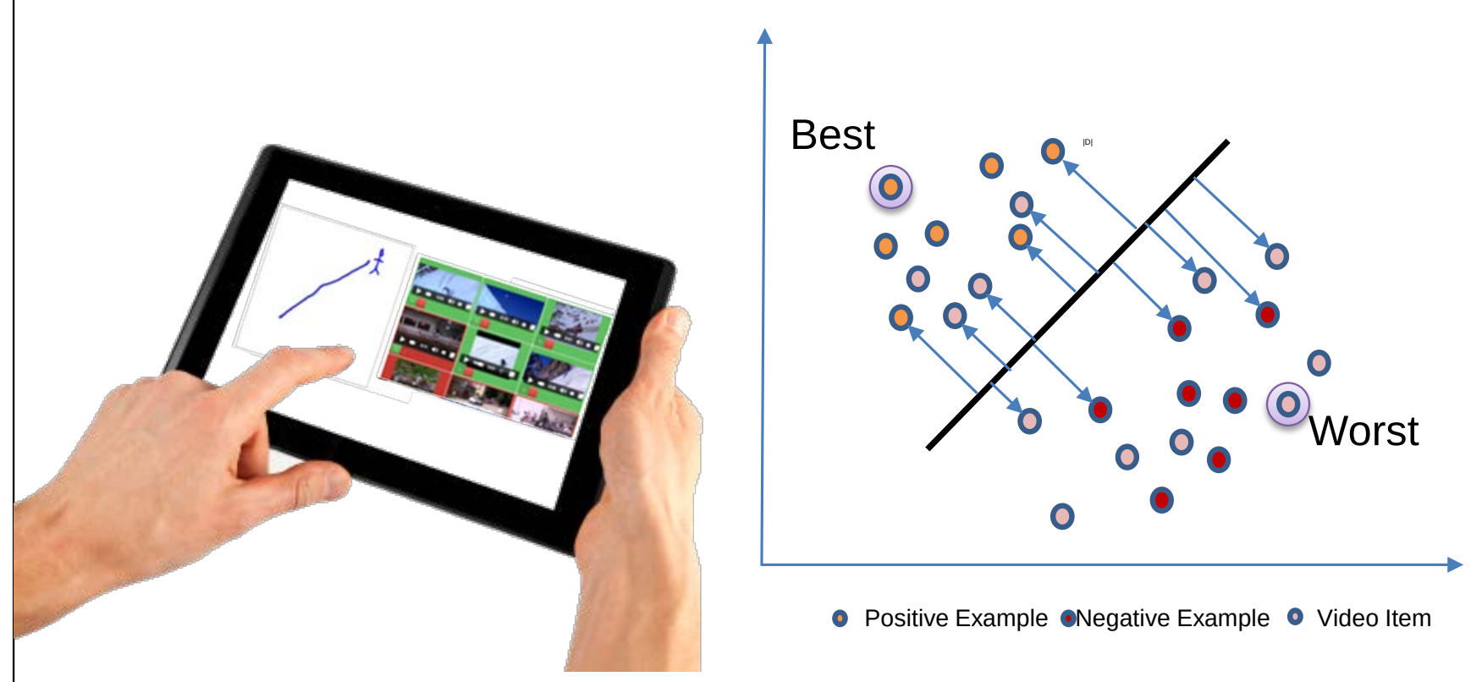
A spatio-temporal descriptor for representing and indexing a video repository for sketch based video retrieval (SBVR). Our descriptor encodes the semantic class, appearance (shape and colour), and the motion direction of each foreground object within the video. The appearance of the background is also captured. Video clips are described in terms of a descriptor of multiple-facets:



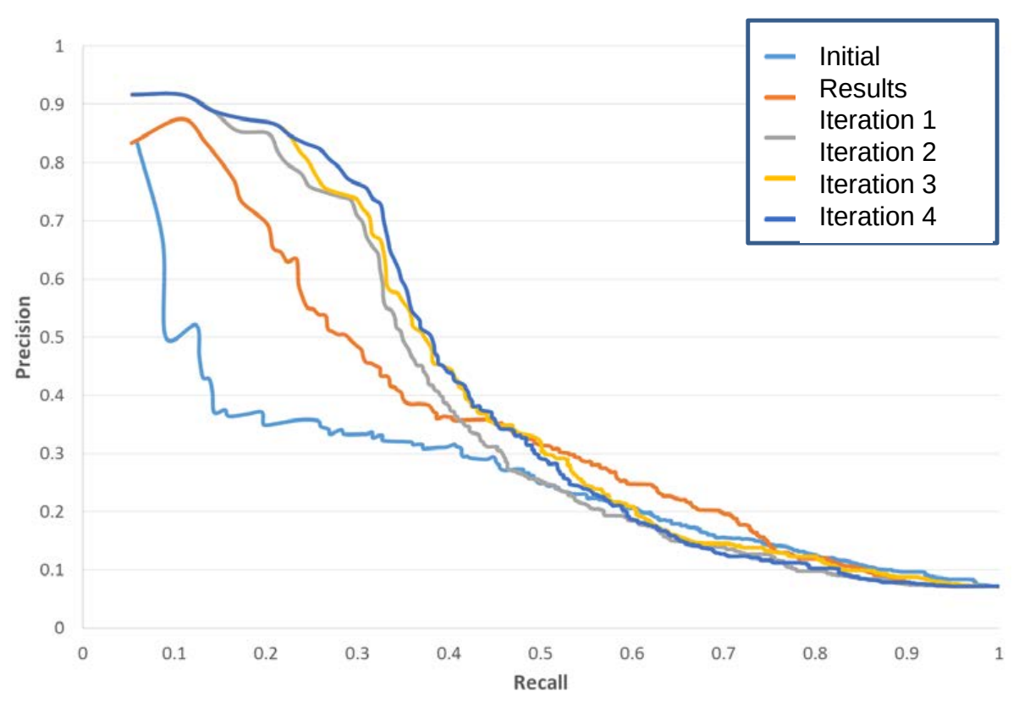
The combination of sketch and text within a query, coupled with a fast video indexing scheme, results in a 'hybrid' SBVR system capable of searching 700 videos in less than one second; several orders of magnitude faster than prior hybrid SBVR systems [6]. Matching based on descriptor achieves an **MAP of 32%**.

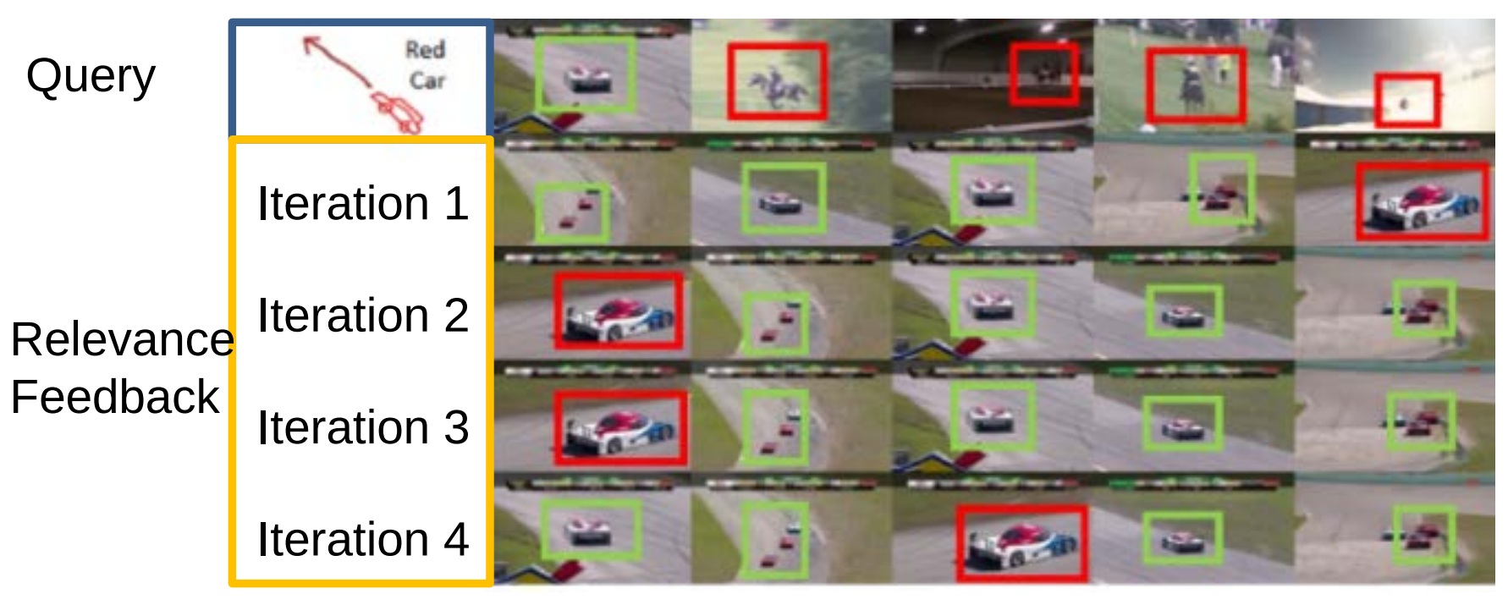


Interactive feedback allows effective improvement to State-of-the-Art



The significant performance benefits of an index-based matching approach for SBVR are near-instantaneous full database search over hundreds of videos. This raises the opportunity of working with the user "in the loop" to interactively refine results. Using an ensemble of SVMs results can be re-ranked based on positive and negative feedback. Improving the **MAP by 42%, 46%, 49%, 50%** for iterations 1,2,3,4 respectively





Sketch based Human Pose Retrieval [3,4]

Silhouettes of video frames are extracted and described using a HoG style descriptor. Alternatively sketch queries are parsed into an articulated skeleton and described through joint angles.

We learn a non-parametric mapping between the query space (S) and pose descriptor space (D), using a set of around 230 manually marked up training poses. Valid poses lie upon manifolds in both spaces, each of which is sampled by the training process. A graph-based strategy is used to compute similarity between a query and candidate video frame(pose) by approximating geodesic distance in piecewise-linear manner across these manifold. This similarity score is used to rank each video frame in the video database for relevance to a given query.

Learning Domain Transfer

We manually annotate a set of key poses with a sketch. For a provided query sketch $(q \in S)$ we measure similarity between that sketch and training sketches. We map the distances through a Gaussian:

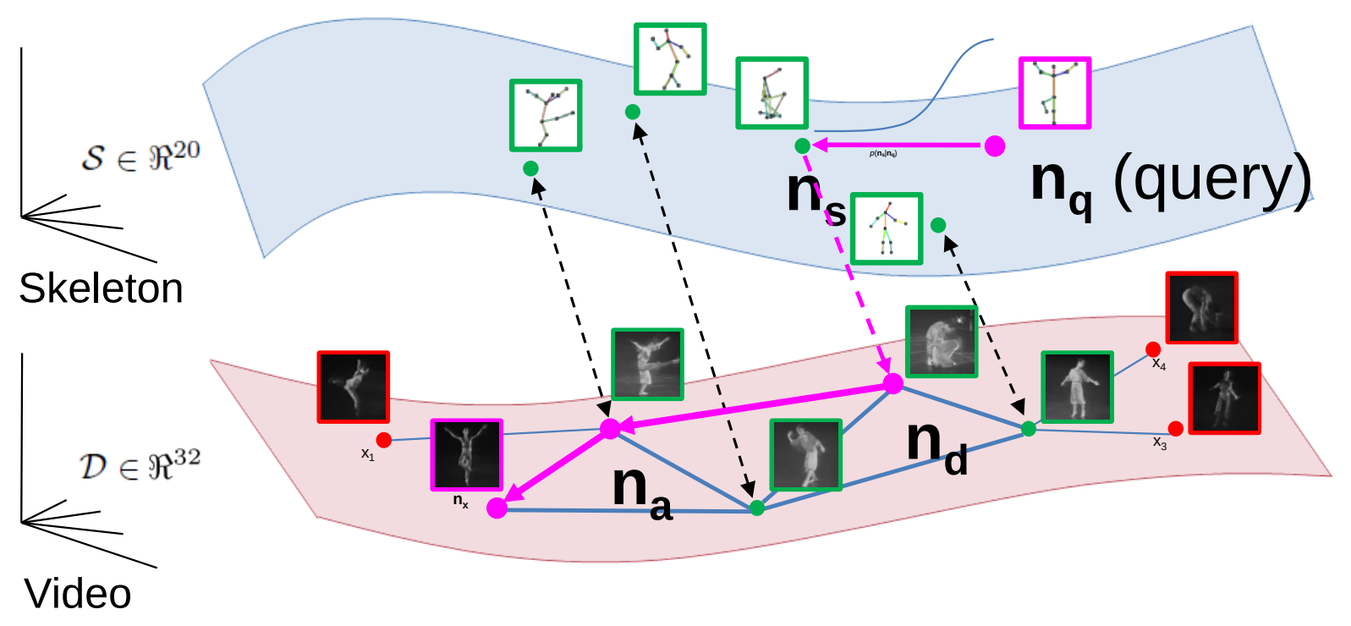
$$p(s|q) \propto exp - \frac{|q-s|^2}{2\sigma}.$$

With the mapping of $s \mapsto d$ from annotations, we can compute the shortest distance across $\mathcal{G}$ to any other node i.e. frame.

$$p(n_x|n_d) = \prod_{\{a,b\}\in\mathcal{G}} 1 - p(n_a|n_b).$$

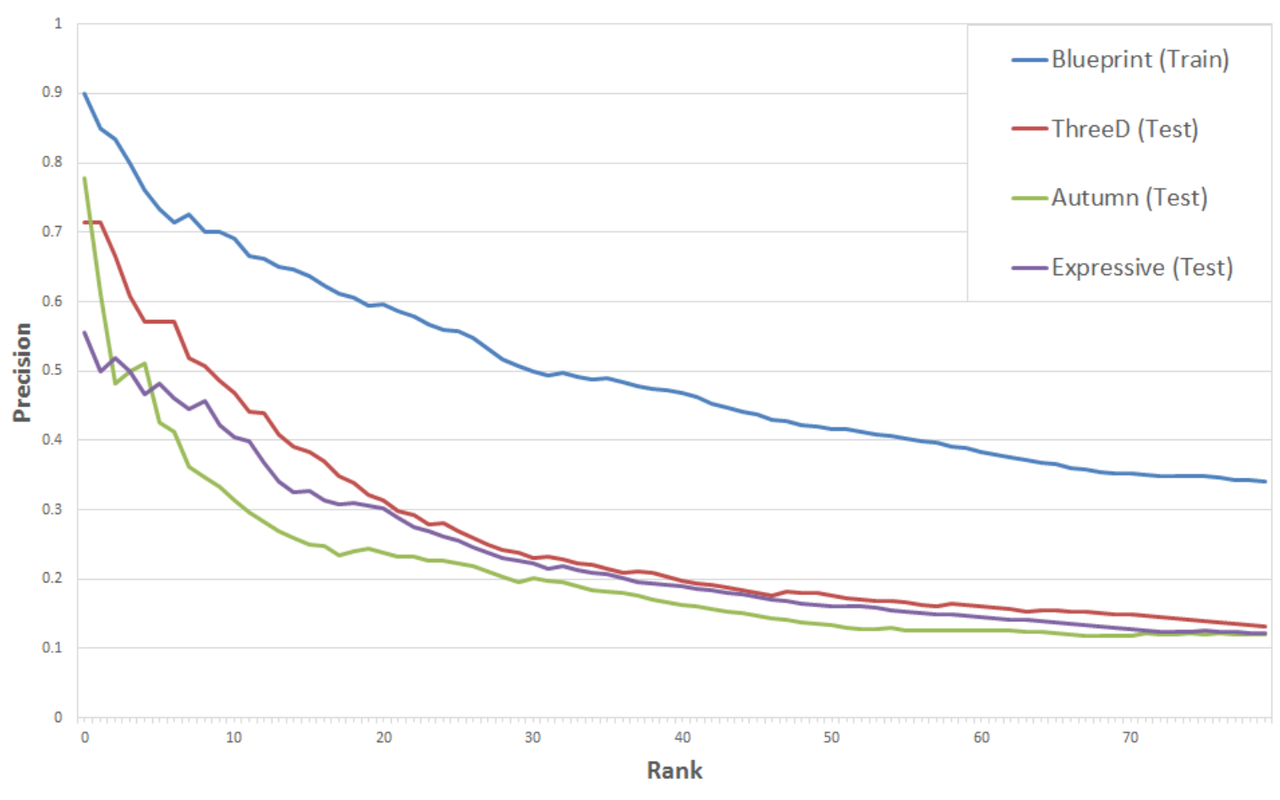
Combining these equations, we compute the joint probability of any video frame n_x in our dataset.

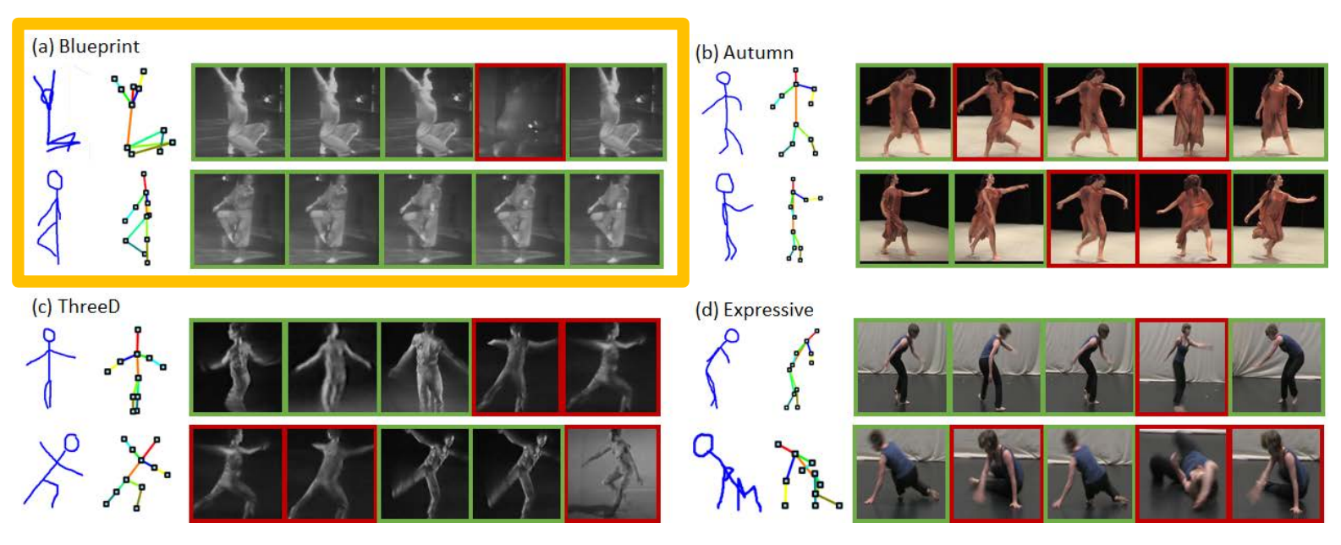
$$p(n_x|q) = p(s|q)p(n_d|n_x).$$



$p(\mathbf{n_x}|\mathbf{n_d}) = p(\mathbf{n_x}|\mathbf{n_a})\, p(\mathbf{n_a}|\mathbf{n_d})$

$p(\mathbf{n_x}|\mathbf{n_q}) = p(\mathbf{n_x}|\mathbf{n_a})\, p(\mathbf{n_a}|\mathbf{n_d})\, p(\mathbf{n_s}|\mathbf{n_q})$

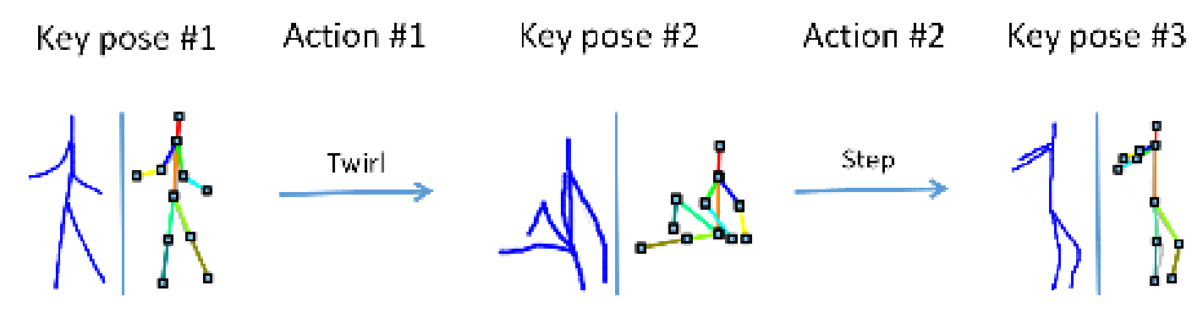




Train MAP: Blueprint **60%**
Test MAP:
ThreeD **32%**, Autumn **48%**, Expressive **38%**

Sketched Visual Narratives [4,5]

We propose a graph-based representation for sketched Visual Narratives that describe events comprising more than one action based on the concept of "Motion Graphs" [7].



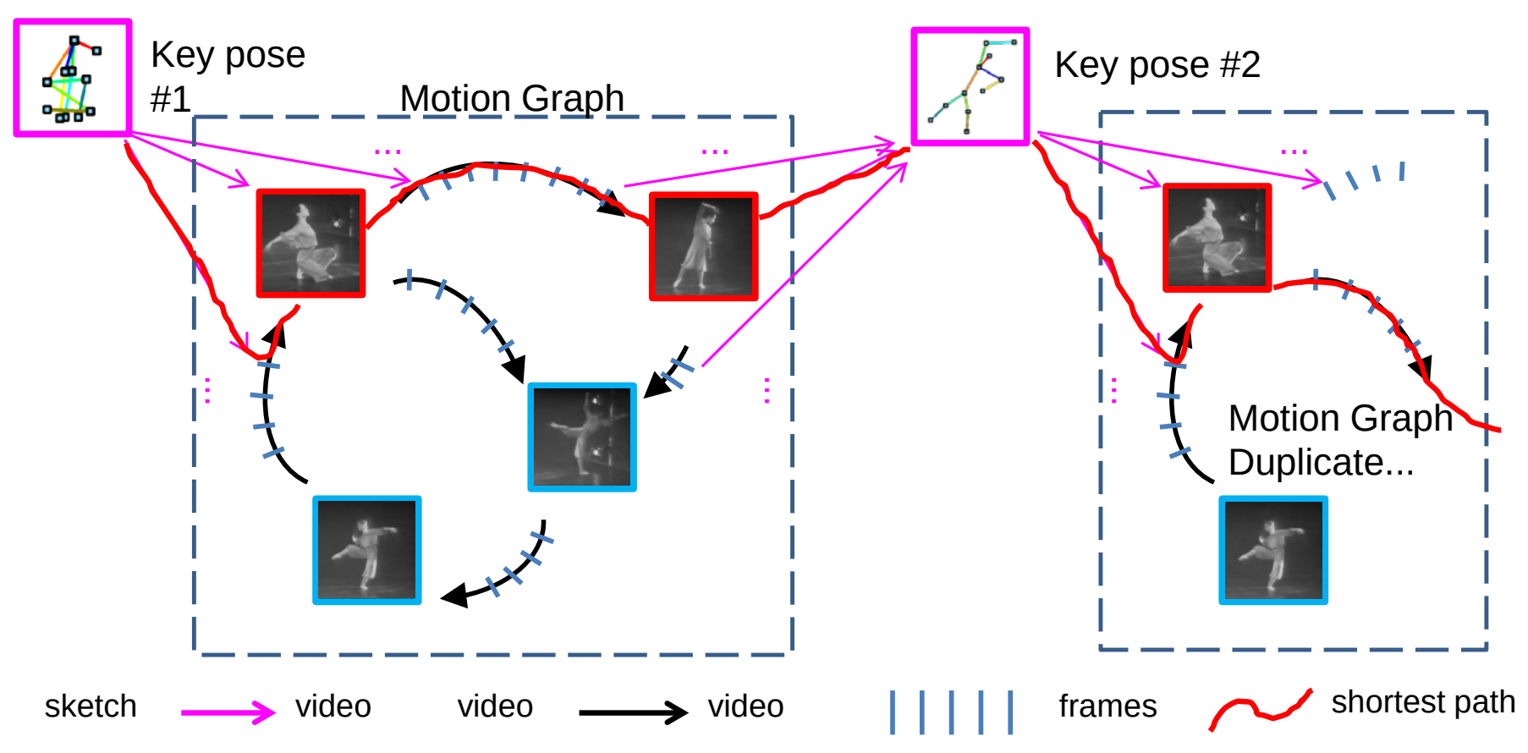
Extending the pose retrieval system, we explore the application of our graph representation. We demonstrate the ability to retrieve and synthesise using a sequence of poses interspersed by actions linking those poses.

Video Synthesis

A directed graph is constructed from video fragments comprising sequential blocks of frames (blue marks on edges) linked seamlessly at transition frames (blue nodes). Sketched key poses (magenta nodes) are added as virtual nodes linking copies of the motion graph. The path with lowest cost (red), by:

$$C = w_p C_{Target} + w_a C_{Action} + w_t C_{Time}.$$

from first to last key pose yields the new choreographic sequence. Where the components of the cost are defined as:



Target C_{Target} $p(n_x|q) = p(s|q)p(n_d|n_x).$

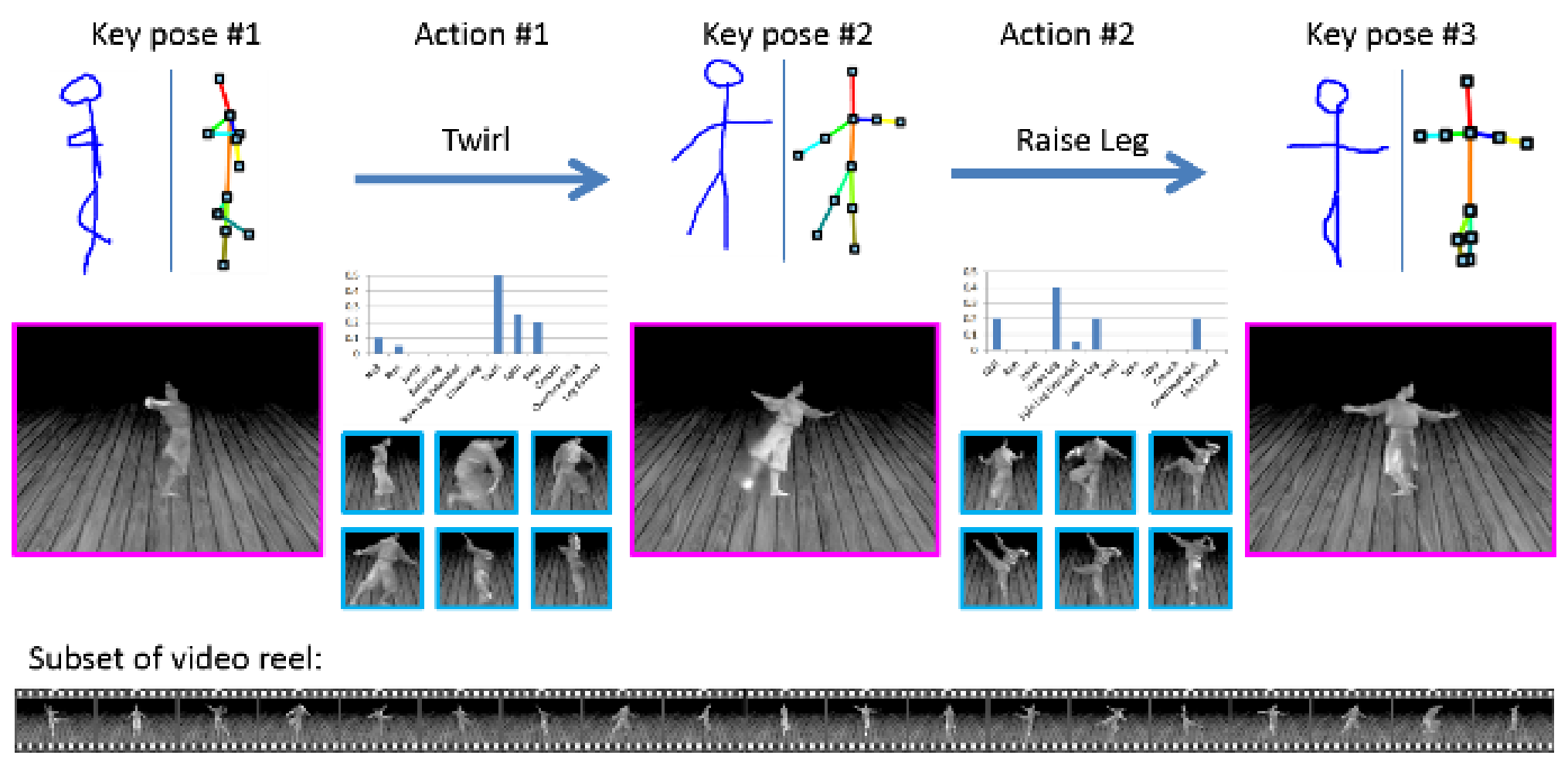
The target difference between the query pose and the frame using the distance from the pose retrieval.

Action $C_{Action} = \frac{1}{k-1}\sum_{l=1}^{k-1}|A(l_i) - A(q_i)|\,.$

Sliding window SVM trained for gesture recognition (black box)

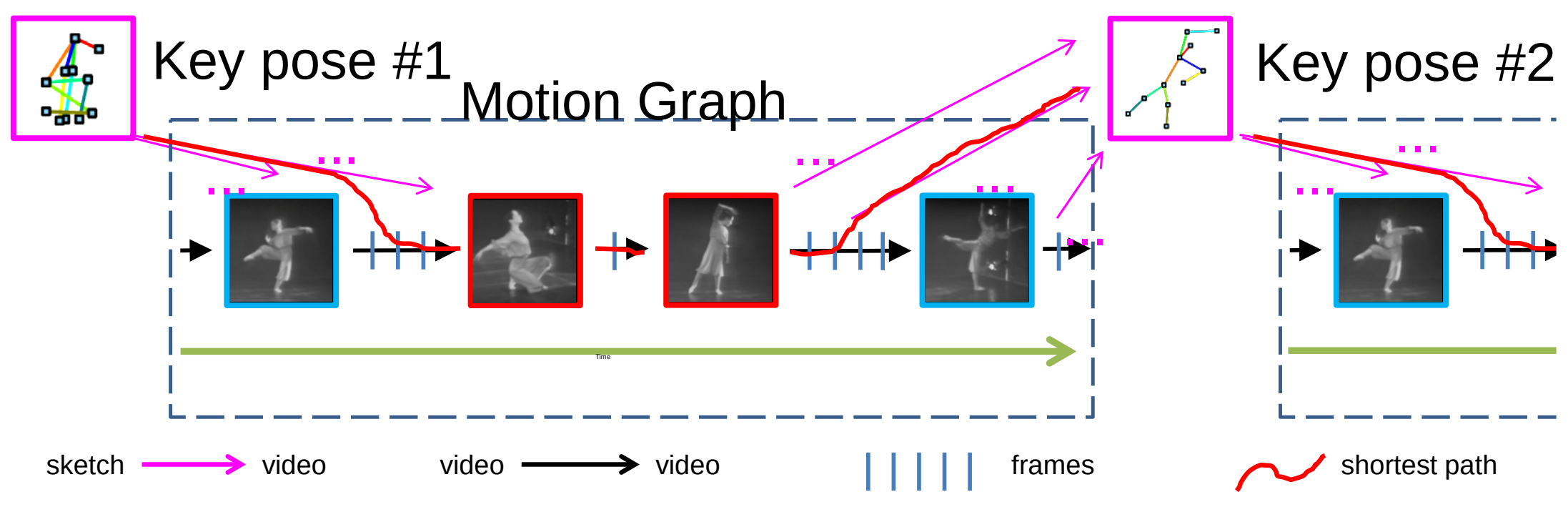
Time $C_{Time} = S(\sum_{i=1}^{k-1}||l_i| - L|)$

Penalise deviation from an idealised duration (user specified with action labels).



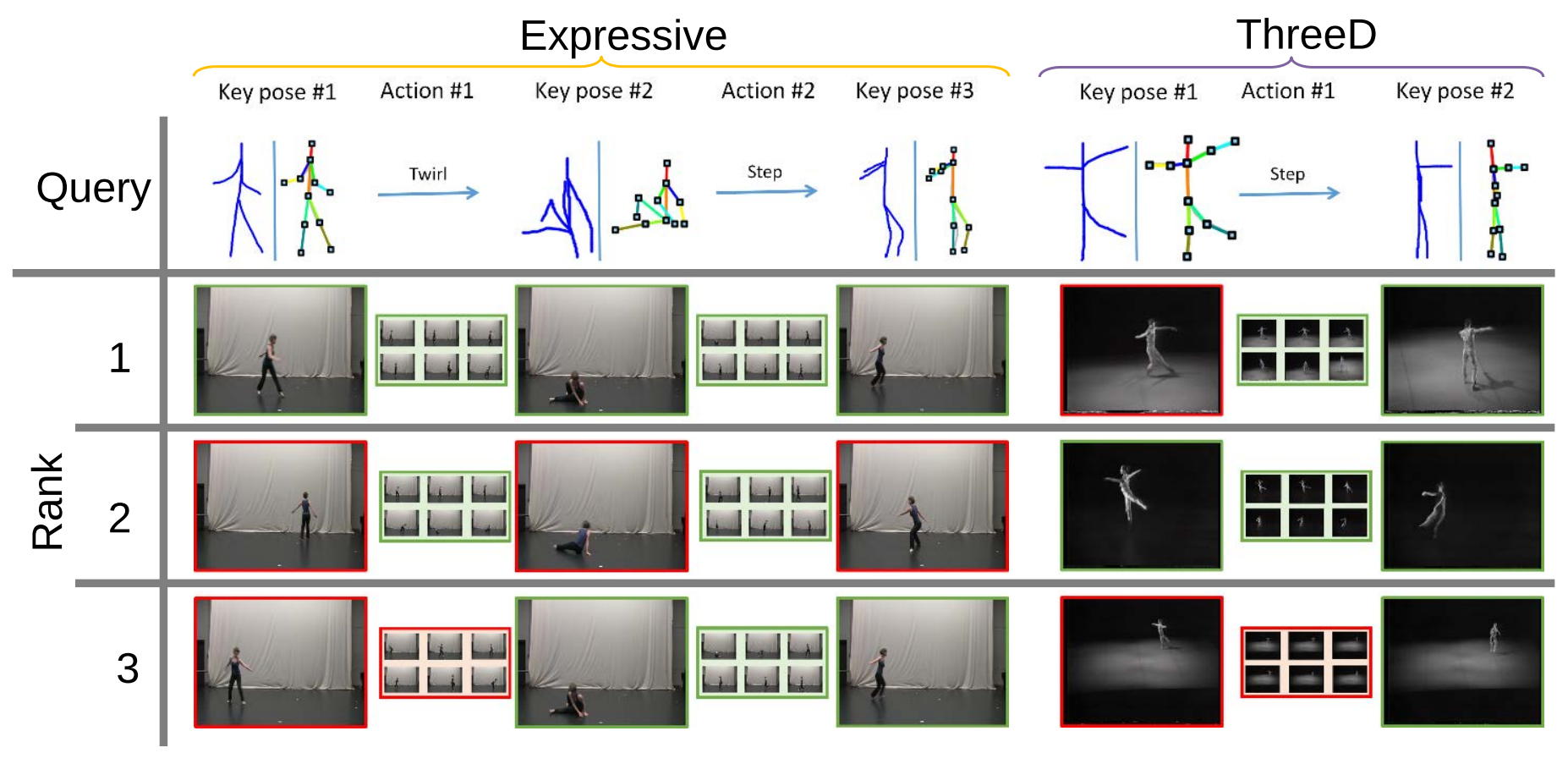
Video Retrieval

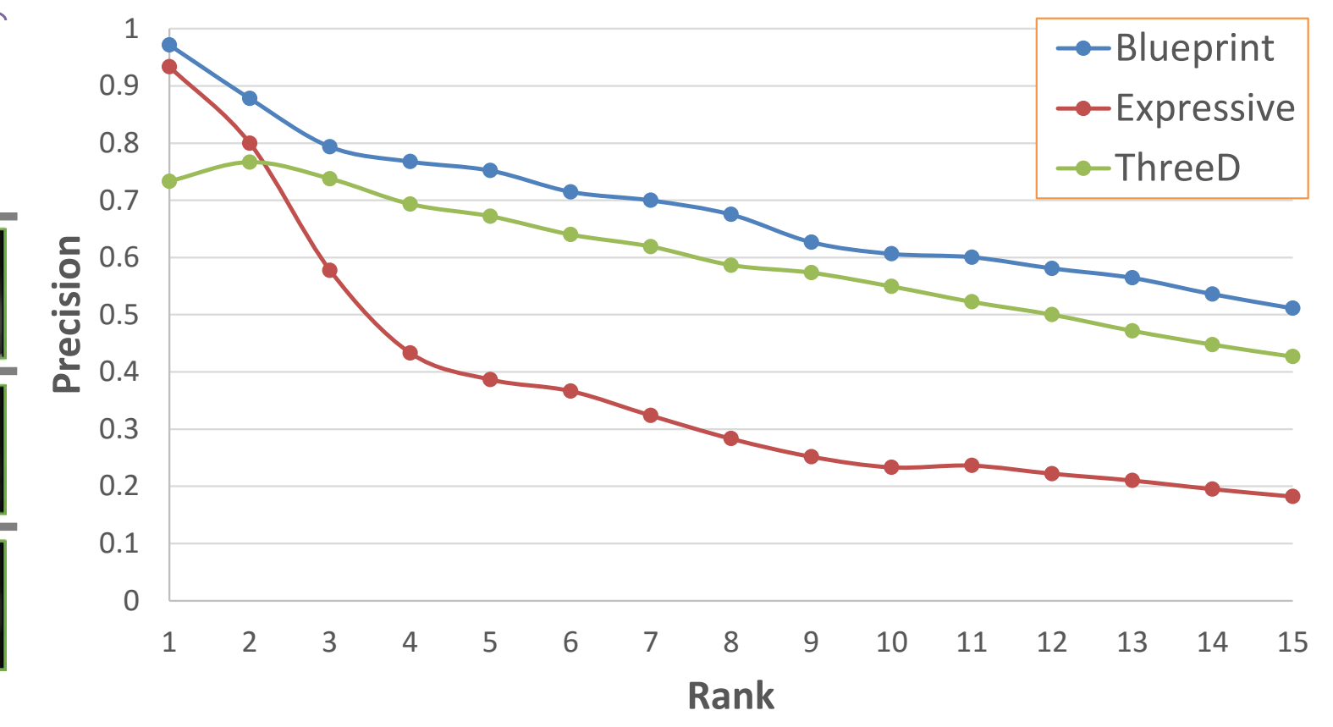
Video Search via the graph representation. Similarly to the synthesis a directed graph is constructed from the video comprising sequential frames (blue marks on edges) linked at salient instants (blue nodes). Sketched key poses (magenta nodes) are added as virtual nodes that connect duplicate copies of the motion graph. The virtual nodes are linked to the nodes representing salient frames. The path across the graph -- from first virtual source node to last virtual sink node -- with lowest cost (red), by eq. 5.1, from first to last key pose yields the most relevant video sub-sequence.



The retrieval of video segments results in an MAP of:

Blueprint **77% MAP**
Expressive **75% MAP**
ThreeD **71% MAP**





References

[1] S James and J Collomosse. "Interactive Video Asset Retrieval using Sketched Queries". Proceedings of Conference on Visual Media Production. ACM. 2014.

[2] S James and J Collomosse."Annotated Sketches for Intuitive Video Retrieval". BMVA/AVA Workshop Biological and Machine Vision. Perception Journal. 2011.

[3] M Fonseca and S James and J Collomosse. "Skeletons from Sketches of Dancing Poses". Proceedings VL/HCC 2012. IEEE. 2012

[4] S James M Fonseca and J Collomosse. "ReEnact: Sketch based Choreographic Design from Archival Dance Footage". Proceedings of ACM International Conference on Multimedia Retrieval (ICMR). ACM. 2014.

[5] S James. "Visual Narratives: Free-hand Sketch for Visual Search and Navigation of Video". PhD Thesis. University of Surrey. 2015.

[6] R Hu and S James T Wang and J Collomosse. "Markov Random Fields for Sketch based Video Retrieval". Proceedings ACM International Conference on Multimedia Retrieval (ICMR). ACM. 2013.

[7] L Kovar, M Gleicher, F Pighin. "Motion Graphs ". ACM Transactions on Graphics (TOG). 2002

Acknowledgements

Attendance supported by:
University College London
and
Engineering and Physical Sciences Research Council (grant EP/K023578/1)