Research Statement – Subhabrata Bhattacharya

My research in computer vision brings together machine learning, insights from psychology, computer graphics, algorithms, and a great deal of computation. Over the last few years, I have had the opportunity to explore several broad areas of research in Computer Vision - including:

Recognizing Complex Events in Consumer Videos

The goal of complex event recognition [9–12] is to automatically detect high-level events in a given video sequence. However, due to the fast growing popularity of such videos, especially on the Web, solutions to this problem are in high demand. A feasible solution can directly make video search and retrieval more efficient and rewarding experience for the users. This can also help track user interest based on the video contents they watch, thereby promoting advertisement of certain products. Furthermore, it can help broadcast agencies predict important statistics about a video such as virality of views, geographical location of viewers etc. moments after a video is uploaded, so that channel bandwidth could be optimized. In addition, such systems can provide human observers with meaningful textual recounting of a video in a relatively short time without substantial human intervention. That said, this in itself is an extremely challenging problem and requires thorough algorithmic breakthroughs at multiple tiers. My research attempts to address some of the sub-problems which are crucial in context of complex event recognition and are listed as follows:

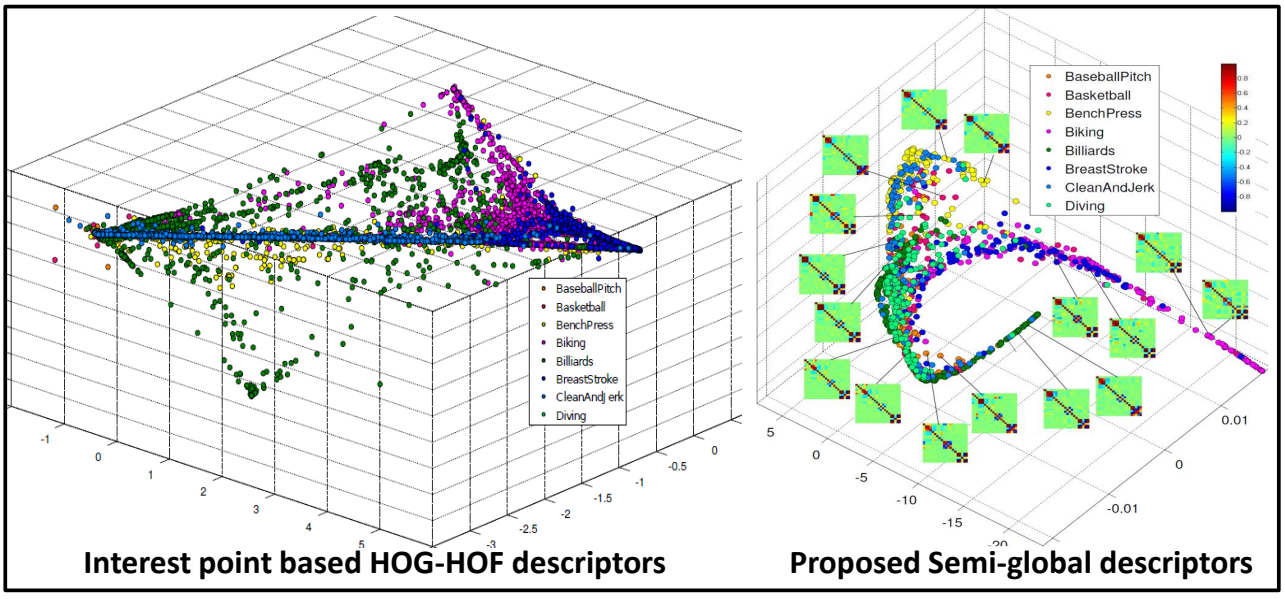

Figure 1: Each circle represents a lower-dimensional manifestation of descriptors from video samples in UCF50 human actions dataset (8 classes). Both types of descriptors are mapped to 3 dimesional space for ease of visualization. Sample covariance matrices are shown as insets to some circular dots. Note how our descriptors form relatively clear cluster boundaries.

(a) Design of features: Within the purview of this effort, we explore two complementary sources of information to design features that are useful for content based video analysis in realistic scenarios. The first one is semi-global in nature, computed from small segments from the video [5], while the second one is based on ambient camera motion [3] present during the video capture process.

The **semi-global clip-level descriptor** is a concise representation of a temporal window/clip of subsequent frames from a video rather than localized spatio-temporal patches, which eliminates the use of specific detectors. The descriptor is based on covariance of complementary low-level motion (optical flow and their derivatives, vorticity, divergence etc.) and appearance cues (first and second order derivatives of pixel intensities etc.). Since covariance matrices capture joint statistics between individual low-level feature modalities, they automatically transform our random vector of samples into statistically uncorrelated random variables, leading to a compact representation of a video. Fig. 1 provides an insight on the discriminative capability of both the HOG-HOF based descriptors and the proposed covariance matrix based descriptors.

In addition to the descriptor itself, we investigate two sparse coding based approaches [5] to use the descriptor in context of action and gesture recognition. Within this, the sparse approximation of a set of covariance matrices is treated as a

determinant maximization problem, where the bases (covariance matrices) are obtained from training videos. We compare this approach with a sparse linear approximation alternative suitable for equivalent vector spaces of covariance matrices using Orthogonal Matching Pursuit. We show the applicability of our video descriptor and the associated recognition algorithms through various experiments on challenging datasets. Our experiments provide promising insights in large scale video analysis.

Camera-motion is often an under-exploited cue when it comes to the analysis of videos depicting complex events in consumer uploaded videos. Complex events like "Attempting a board trick" and "Parkour" usually have a lot of jittery camera motion coupled with pan and tilt motions. Similarly, videos depicting events such as "Wedding Ceremony" and "Birthday Party" are mostly captured by stationary cameras with limited pan and some amount of zoom. The objective of this effort [3] is to investigate an efficient set of methodologies, that can be leveraged to represent videos in terms of their ambient camera motion in large scale, without resorting to computationally prohibitive full-3D reconstruction techniques.

We devise this novel representation on top of inter-frame homographies which serve as coarse indicators of the camera motion. Next, using Lie algebra of projective groups, we transform the homography matrices to an intermediate vector space that preserves the intrinsic geometric structure of the transformation (Fig. 2). Multiple time series are then constructed from these mappings. We perform an exhaustive analysis of effective features that can be computed from these time-series based on theoretical foundations from both linear (Hankel matrices) and non-linear (Chaotic invariants) dynamical systems. Features computed on these time series are used for discriminative classification of video shots. Our proposed camera motion based shot classification outperforms previously published algorithms and achieves comparable performance to an implementation that involves recovery of structure from motion on our dataset of eight shot categories. This encourages us to evaluate our method for complex event recognition in challenging datasets [9, 10], which demonstrates conclusive evidence towards its applicability in open-source video analysis.

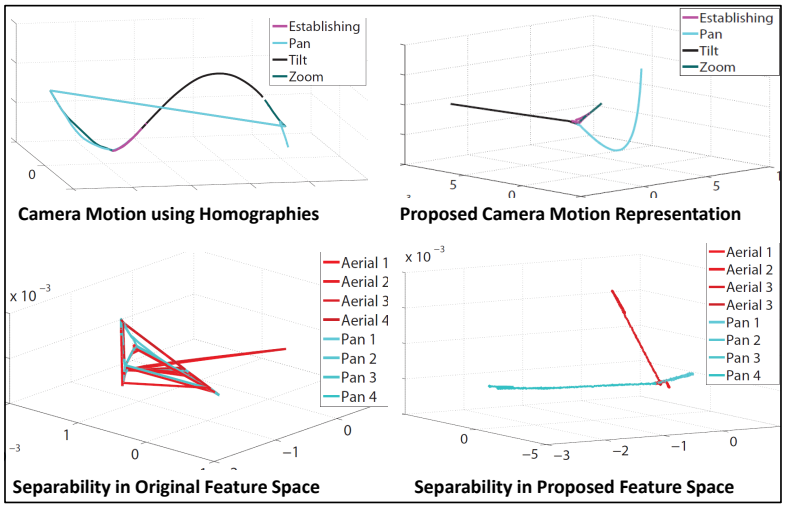

Figure 2: Discriminativity of our proposed representation of shots in contrast to pure frame-to-frame homographies. Figures on top show four classes of shots (Establishing, pan, tilt and zoom) in both feature space. Bottom figures show clear separability of "aerial" and "pan" shot classes in the proposed feature space.

(b) Engineering computationally efficient intermediate representations: Designing intermediate representations on top raw features is very crucial for any recognition algorithm in order to handle outliers efficiently and reduce processing of large volumes of high dimensional data. A popular approach in this context is the Bag-of-Visual-Words (BoVW) methods where raw features extracted in a video or image are quantized using common clustering algorithms and reduced to a histogram representation, which becomes the intermediate representation or signature for a video or image. We present an efficient alternative [6] to the traditional vocabulary based on BoVW methods used for visual classification tasks.

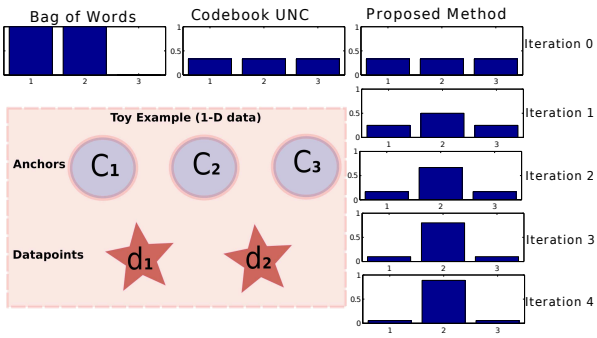

Figure 3: Toy example contrasting the proposed representation against traditional BoW and soft-assignment BoW. Note that the proposed representation is initially identical to soft BoW but diverges since it maximizes an instance-level likelihood score.

Our representation (Fig. 3) is both conceptually and computationally superior to the bag-of-visual words: (1) We iteratively generate a **Maximum Likelihood estimate** of an instance given a set of characteristic features in contrast to the BoVW methods (2) We randomly sample a set of characteristic features called **anchors** instead of employing computation intensive clustering algorithms used during the vocabulary generation step of BoVW methods. Our performance compares favorably to the state-of-the-art on experiments over three challenging human action and a scene categorization dataset, demonstrating the universal applicability of our method.

We integrate the above representation scheme to detect semantically accurate, human-understandable mid-level spatio-temporal concepts for modeling complex events. To this we introduce a benchmark dataset for spatio-temporal concepts extracted from amateur videos depicting complex events. This dataset consists of 104 mutually exclusive, concept categories over 10,000 annotated audio visual samples extracted from NIST's TRECVID MED 2011 event corpus that replicates com-

plex events observed in common video footages. Detectors are trained on the proposed anchors based representation specific to each concept category on different information modalities (motion, static, and audio). This approach achieved respectable target detection [10] in the annual NIST TRECVID Multimedia Event Detection 2011 competition.

(c) Formulating complex event models: Just as low-level features and the associated intermediate representations are crucial for recognition, efficient complex event models can be created if temporal dynamics are exploited effectively exploited. So far researchers have proposed the use of various configurations of graphical models in this context. Although these models are mathematically intuitive and elegant, they are computationally complex and require extensive training coupled with substantial domain knowledge.

Here we represent each video depicting a complex event, as an ordered vector time-series, where each time-step is a vector containing confidences returned by a set of pre-trained spatio-temporal concept detectors [10]. Using, foundations from linear dynamical systems, we extract two complementary features, the first is based on Block Hankel matrices, which captures dependencies between each observation vector, within the context of the entire time-series. The second exploits statistically meaningful characteristics from multiple interacting time-series such as lag-independence, harmonics, frequency proximity etc. We also integrate the above feature computation steps into a Bayesian concept selection framework, that automatically identifies the concepts necessary to achieve a respectable trade-off between accuracy and computational efficiency of the recognition process. Experiments conducted on NIST's, TRECVID datasets for Multimedia Event Detection (MED 2011 & MED 2012), demonstrate how our proposed method [2] outperforms the state of the art in context of complex event recognition.

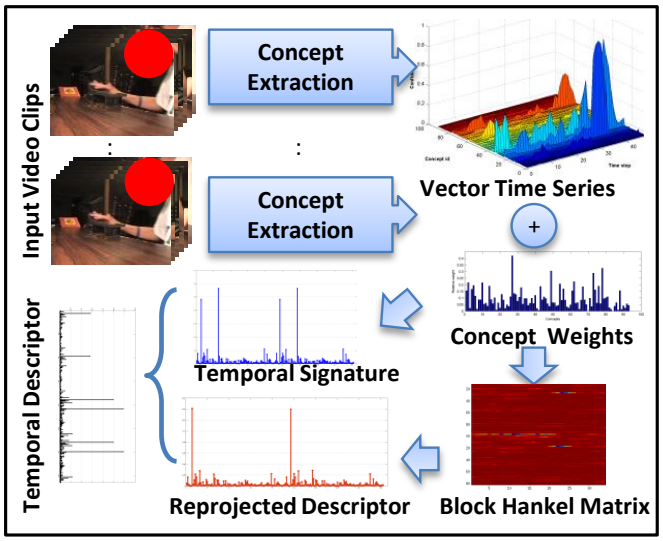

Figure 4: Proposed temporal feature extraction from a typical video. We build our complex event recognition computational pipeline based on this methodology.

Computational Photo-aesthetics

The deluge of image hosting Web sites and increasing affordability of consumer grade digital cameras, have introduced two new problems in image sharing perspective: the first is the ability to select the best-looking ones from a large pool of photographs captured during certain occasion. The next is the flexibility to edit a photograph with minimal photographic compositional knowledge so that the result looks reasonably better than the original ones. These two key issues motivate us to propose a set of novel algorithms [7, 8] that enable naive users to improve the visual aesthetics of their digital photographs using several novel spatial recompositing techniques. This work differs from earlier efforts in two important aspects: (1) it focuses on both photo quality assessment and improvement in an integrated fashion, (2) it enables the user to make informed decisions about improving the composition of a photograph.

Figure 5: Photo-quality enhancement: The images in the left are input to our composition enhancement tool, while their enhanced counterparts are shown in right.

The tool facilitates interactive selection of one or more than one foreground objects present in a given composition, and the system presents recommendations for where it can be relocated in a manner that optimizes a learned aesthetic metric while obeying semantic constraints. For photographic compositions that lack a distinct foreground object, the tool provides the user with crop or expansion recommendations that improve the aesthetic appeal by equalizing the distribution of visual weights between semantically different regions. The recomposition techniques presented here emphasize learning support vector regression models that capture visual aesthetics from user data and seek to optimize this metric iteratively to increase the image appeal. The tool demonstrates promising aesthetic assessment and enhancement results on variety of images and provides insightful directions towards future research. This work [7] was also nominated for **best paper** in ACM MM 2010 full paper track, which was later extended in [8].

Aerial Video Analysis

Quadrotor helicopters have gained immense visibility in the area of aerial surveillance and reconnaissance over the last decade. Due to their portability, ease of control, low risk of operation and affordable cost of deployment, these low flying platforms are getting popular across law enforcement departments around the world for applications such as tracking vehicles or monitoring suspicious activities. We introduced a technique to solve the problem of tracking objects persistently from surveillance platforms integrating quad-rotor aerial (moving) and ground (fixed) platforms in typical urban scenarios as shown in Fig. 6. Under this framework [4] we track moving objects from a moving aerial platform using a three staged conventional technique [1] consisting of ego-motion compensation, blob detection, and blob tracking with near-realtime precision. A hierarchical robust background subtraction followed by a motion correspondence algorithm is applied to track objects from the ground surveillance camera.

We further refine [13] the metadata available at the airborne camera and along with the calibration parameters of the ground camera, we are able to transform the objects position in both cameras local coordinate system to a generic world coordinate system. Trajectories obtained in terms of the world coordinates are then merged assuming temporal continuity. False candidate trajectories are eliminated using similarity metric based on color intensity of the object that generated it. Our system has been tested in 3 real-world scenarios where it has been able to merge trajectories successfully in 80% of the cases. The tools developed [1, 13] as part of this project were important contributions towards UCF–Lockheed Martins involvement in the **DARPA Video Image Retrieval and Analysis Tool (VIRAT)** program and is extensively used to extract motion- compensated chips depicting human activities from aerial videos.

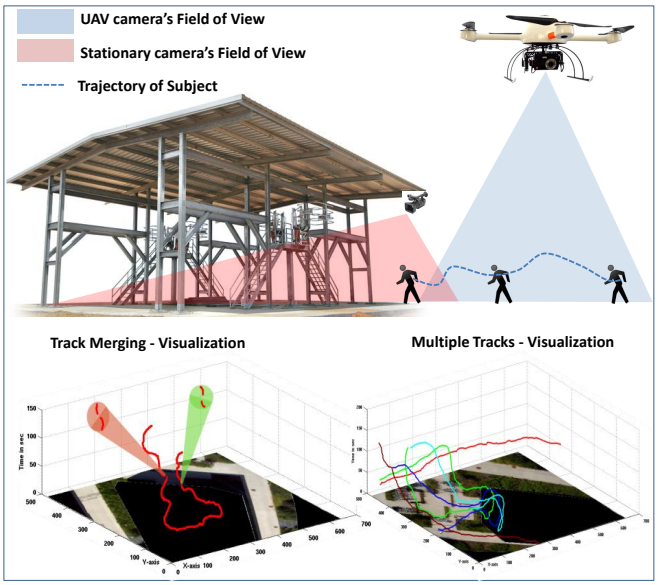

Figure 6: (Top) A typical urban surveillance scenario, (Bottom) shows results from our track merging approach.

Collaborations and Outreach

One of the advantages of working in such a vibrant field is the opportunity for fruitful collaboration across both industries and academia. Currently In past, I have collaborated with researchers at Columbia University [11, 12], Carnegie Melon University [10], University of Michigan [6] and University of Klagenfuert [13]. I have been fortunate to publish with several renowned researchers in computer vision, and participated in research projects with industrial partners such as Lockheed Martin [1], SRI Sarnoff, Google Research. I have also interned in two separate occasions with Microsoft Research and Intel Labs [7, 8] during the summers of 2012 and 2010, respectively. Having worked for research and development in systems (IBM Systems & Tech. Groups and Infosys Tech. Ltd.) provides me with a natural edge to effectively contribute to large groups.

In addition to the two high profile conferences in computer vision and multimedia, I regularly speak at specialized workshops on recognition. My work has been funded by DARPA, IARPA, Intel. I am actively involved in writing grant proposals for AFOSR, NSF and NASA.

References

[1] S. Bhattacharya, H. Idrees, I. Saleemi, S. Ali, and M. Shah. Moving object detection and tracking in forward looking infra-red aerial imagery. *Machine Vision Beyond Visible Spectrum*, pages 221–252, 2011.

[2] S. Bhattacharya, M. Kaleyeh, R. Sukthankar, and M. Shah. Understanding temporal dynamics of low-level concepts for complex event recognition. In *Proc. of ACM Multimedia (under review)*, 2013.

[3] S. Bhattacharya, R. Mehran, R. Sukthankar, and M. Shah. Cinematographic shot classification and its application to complex event recognition. *IEEE Transactions on Multimedia (TMM) (under review)*, 2012.

[4] S. Bhattacharya, M. Quaritsch, B. Rinner, and M. Shah. Quadrotors for persistent urban surveillance: A case study. *UCF Technical Report (TR-CAM-1901)*, Dec. 2009.

[5] S. Bhattacharya, N. Souly, and M. Shah. Covariance of motion and appearance features for spatio temporal recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (under review)*, 2012.

[6] S. Bhattacharya, R. Sukthankar, R. Jin, and M. Shah. A probabilistic representation for efficient large scale visual recognition tasks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2593–2600, 2011.

[7] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM Multimedia (MM)*, pages 271–280, 2010.

[8] S. Bhattacharya, R. Sukthankar, and M. Shah. A holistic approach to aesthetic enhancement of photographs. *Transactions of Multimedia Computing, Communications and Applications (TOMCCAP)*, 7(Supplement):21, 2011.

[9] H. Cheng, J. Liu, S. Ali, O. Javed, Q. Yu, A. Tamrakar, A. Divakaran, H. S. Sawhney, R. Manmatha, J. Allan, A. Hauptmann, M. Shah, S. Bhattacharya, A. Dehghan, G. Friedland, B. M. Elizalde, T. Darrell, , M. Witbrock, and J. Curtis. Sri-sarnoff aurora system at trecvid 2012: Multimedia event detection and recounting. *Proc. of NIST TRECVID and Workshop*, Dec. 2012.

[10] H. Cheng, A. Tamrakar, S. Ali, Q. Yu, O. Javed, J. Liu, A. Divakaran, H. S. Sawhney, A. Hauptmann, M. Shah, S. Bhattacharya, M. Witbrock, J. Curtis, G. Friedland, R. Mertens, T. Darrell, R. Manmatha, and J. Allan. Team sri-sarnoffs aurora system@ trecvid 2011. *Proc. of NIST TRECVID and Workshop*, Dec. 2011.

[11] Y. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. *Proc. of NIST TRECVID Workshop*, Dec. 2010.

[12] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval (IJMIR)*, November 2012.

[13] M. Quaritsch, K. Kruggl, D. Wischounig-Strucl, S. Bhattacharya, M. Shah, and B. Rinner. Networked uavs as aerial sensor network for disaster management applications. *Elektrotechnik und Informationstechnik (E&I)*, 127(3):56–63, 2010.