Recognition of Complex Events in Open-Source Web-Scale Videos: A Bottom up approach

Subhabrata Bhattacharya
Center for Research in Computer Vision
University of Central Florida
Orlando, FL
subh@cs.ucf.edu

ABSTRACT

Recognition of complex events in unconstrained Internet videos is a challenging research problem. In this symposium proposal, we present a systematic decomposition of complex events into hierarchical components and make an in-depth analysis of how existing research are being used to cater to various levels of this hierarchy. We also identify three key stages where we make novel contributions which are necessary to not only improve the overall recognition performance, but also develop richer understanding of these events. At the lowest level, our contributions include (a) compact covariance descriptors of appearance and motion features used in sparse coding framework to recognize realistic actions and gestures, and (b) a Lie-algebra based representation of dominant camera motion present in video shots which can be used as a complementary feature for video analysis. In the next level, we propose an (c) efficient maximum likelihood estimate based representation from low-level features computed from videos which demonstrates state of the art performance in large scale visual concept detection, and finally, we propose to (d) model temporal interactions between concepts detected in video shots through two new discriminative feature spaces derived from Linear dynamical systems which eventually boosts event recognition performance. In all cases, we conduct thorough experiments to demonstrate promising performance gains over some of the prominent approaches.

Category and Subject Descriptors: H.4 [Information Systems Applications] : Miscellaneous

Keywords: Complex Event recognition, Multimedia Event Detection, Covariance Matrices, Lie Algebra, Riemannian manifolds, Cinematographic Techniques, Shot classification, Video Descriptors, Maximum Likelihood Estimates, Linear Dynamical Systems, Block Hankel matrices

1. INTRODUCTION & MOTIVATION

Hundreds of hours of multimedia content are uploaded in video sharing portals everyday. Most of these videos are captured by amateur users with limited cinematographic knowledge, and are subject to camera motion, background clutter and frequent illumination changes. Usually these videos depict high-level social events - such as a music concert, birthday party or instructional events such as cooking a recipe or teaching a piano lesson. Thus, sifting through such enormous collections for a specific event is a crucial task and is often painstakingly frustrating given the technological maturity of current video browsing algorithms. Most algorithms, rely heavily on the generosity of the uploader to provide meaningful textual labels relevant to the uploaded video content. Since such textual labels are frequently noisy [10, 11], automatic analysis of such videos are gradually attracting a lot of researchers from computer vision and multimedia communities.

One task within the realm of automatic video content analysis is the recognition of complex events contained in the videos. The goal of complex event recognition is to automatically detect high-level events in a given video sequence. In addition to the obvious benefit of making video search and retrieval more efficient and rewarding experience for the user, tracking user interest based on the video content they watch, may (a) **promote effective advertisement** of certain products. Also, such rich automatic semantic description of videos can help broadcast agencies predict important statistics such as virality of views, geographical location of viewers etc. moments after videos are uploaded – thereby (b) **optimizing broadcast channel bandwidth**. Furthermore, it would enable human observers with (c) **semantically rich textual summary of a video** in a relatively short duration without substantial human intervention, for rapid analysis purposes.

2. RECOGNITION OF COMPLEX EVENTS

High-level or complex events are long-term spatio-temporal object interactions that happen under certain scene settings. However, there are several technical challenges involved in understanding complex events from unconstrained videos. Some of them are listed in the following section.

2.1 Technical Challenges

Current approaches rely heavily on classifier-based methods employing directly computable low-level features from videos. Research strongly suggests the joint use of multiple features [6] such as static frame-based features, spatio-temporal features and acoustic features. Since these low-level features are designed with more controlled conditions in mind [2, 9], it is not clearly understood [3–5] if they are capable of capturing discriminative yet relevant information from diverse open-source videos.

Secondly, after features are computed, they are typically quantized into "video words" and each video is reduced to a histogram popularly known as bag-of-X representation (X being a feature modality). Classifiers are trained on these histograms with event labels to obtain models that can be used for testing videos with unknown labels. However, even with promising retrieval results [4, 5],

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://www.sheridanprinting.com/acm/sigmm2/sigmm2.cfm?eid=5E5E01040653.

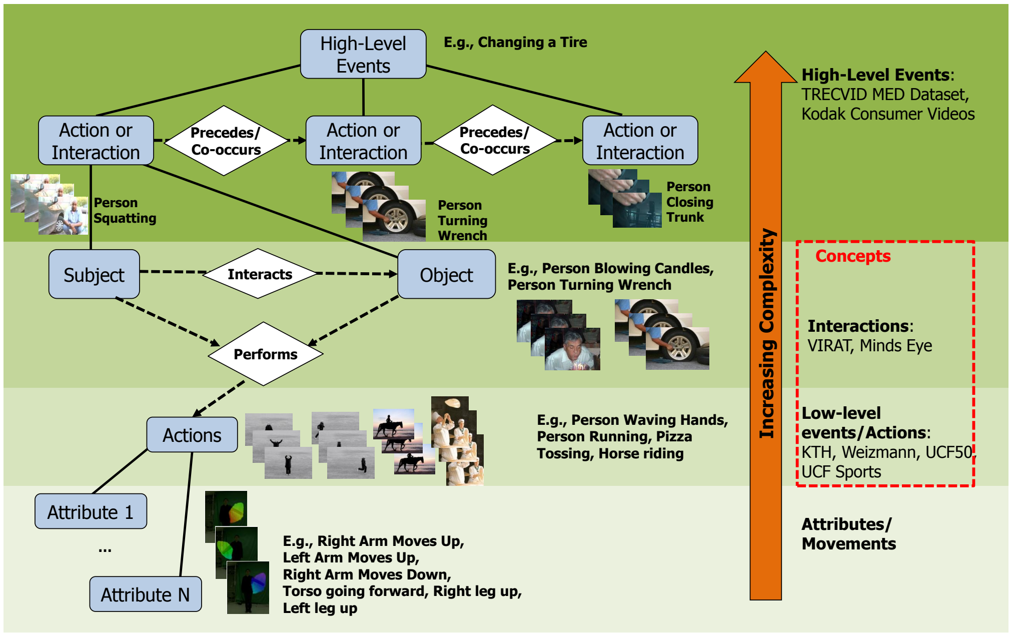

High-Level Events
E.g., Changing a Tire
Action or Interaction
Precedes/ Co-occurs
Action or Interaction
Precedes/ Co-occurs
Action or Interaction
Person Squatting
Person Turning Wrench
Person Closing Trunk
Subject
Interacts
Object
E.g., Person Blowing Candles, Person Turning Wrench
Performs
Actions
E.g., Person Waving Hands, Person Running, Pizza Tossing, Horse riding
Attribute 1
...
Attribute N
E.g., Right Arm Moves Up, Left Arm Moves Up, Right Arm Moves Down, Torso going forward, Right leg up, Left leg up
Increasing Complexity
High-Level Events: TRECVID MED Dataset, Kodak Consumer Videos
Concepts
Interactions: VIRAT, Minds Eye
Low-level events/Actions: KTH, Weizmann, UCF50 UCF Sports
Attributes/ Movements

Figure 1: A hierarchical decomposition as proposed in [6] of complex events, with increased complexity from bottom to top. This bottom-up decomposition helps dividing the original problem of event recognition into tractable and simpler sub-problems with richer semantic understanding.

these methods suffer from the usual disadvantages of quantization used in converting raw features to discrete codewords. Alternatively, X can be replaced by mid-level concepts in the bag-of-X representation to obtain a semantically richer representation of a video [3–6]. A concept can be visualized as a spatial or spatio-temporal entity describing one of the following: (a) object (car, human etc.), (b) scene (forest, beach etc.), (c) action (jumping, running) and, (d) interaction (person riding a bike). Nevertheless these representations are semantically superior, they are unable to capture temporal order between concepts which may be useful for predicting the nature of events and drawing a textual summary of the same (motivation (c) in Section 1).

Thus, to solve this enormous problem, a systematic decomposition of an event is extremely necessary, which not only enables solving simpler and tractable problems but also facilitates developing a better semantic understanding of complex events. We propose the hierarchical decomposition in Fig. 1 appropriate in this context. Through the rest of this proposal, we propose how individual components of the hierarchy in Fig. 1 can be improved to improve the overall performance of recognition of complex events.

2.2 Design of Novel Features

Within the purview of this effort, we explore two complementary sources of information to design features that are fast to compute and also useful for realistic video analysis. Our first feature is computed from covariance of low-level appearance and motion cues obtained from all pixels in a short video clip (typically 20 frames) while the next feature encapsulates ambient camera motion 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.

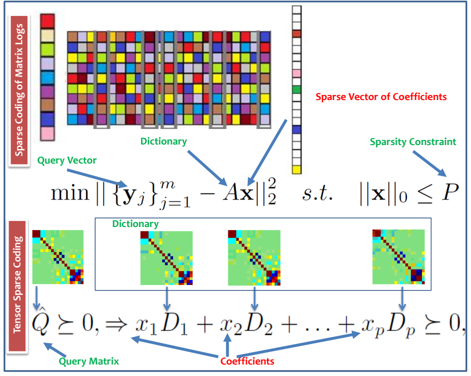

Sparse Coding of Matrix Logs
Sparse Vector of Coefficients
Dictionary
Sparsity Constraint
Query Vector
$\min || \{\mathbf{y}_j\}_{j=1}^{m} - A\mathbf{x}||_2^2 \quad s.t. \quad ||\mathbf{x}||_0 \leq P$
Tensor Sparse Coding
Dictionary
$\hat{Q} \succeq 0, \Rightarrow x_1 D_1 + x_2 D_2 + \ldots + x_p D_p \succeq 0,$
Query Matrix
Coefficients

Figure 2: Two of our proposed sparse coding based classification techniques are illustrated here with the top one using matrix log descriptors as input, while the bottom one operating on covariance matrices directly.

Since these descriptors are computed on a dense temporal sampling basis, most of these can be expressed using only a few discriminative ones to form a dictionary, which can be randomly selected from a collection of labeled videos. This motivates us to resort to sparse coding based classification strategies (Fig. 2) instead of regular SVM classification strategies. Sparse coding on covariance matrices can be nicely formulated as a determinant maximization problem where each query covariance matrix is approximated using a linear combination of dictionary elements. Since vector addition and scalar multiplication of covariance matrices is not closed, additional modifications need to be performed in the original sparse coding formulation which expects a query to be expressed as linear combination of dictionary elements. A schematic illustration of these methods are provided in Fig. 2.

Using our techniques we achieve high recognition rates on the UCF50 [1], and HMDB51 [7]) datasets, which are considered benchmarks for action recognition in unconstrained scenarios. A summary of our results is provided in Tab. 1, where the bottom row

[1] http://server.cs.ucf.edu/~vision/data/UCF50.rar

Method	Desc.	Datasets KTH	 UCF50	 HMDB51
BoVW+SVM	HOG-HOF [8]	92.0%	48.0%	20.2%
BoVW+SVM	COV	81.3%	39.3%	18.4%
SVM	COV	82.4%	40.4%	18.6%
SVM	LCOV	86.2%	47.4%	21.03%
OMP	LCOV	88.2%	53.5%	24.09%
TSC	MAT	93.4%	57.8%	27.16%

Table 1: **Comparison with the state-of-the-art methods: Last two rows show proposed methods against different representations and classification strategies.**

shows two variants of our method, while the top four rows indicate conventional approaches with different feature representations.

Camera-motion is often an under-exploited cue when it comes to the analysis of videos of our concern. 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, tripod-mounted cameras with limited pan and some amount of zoom. The objective of this effort 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.

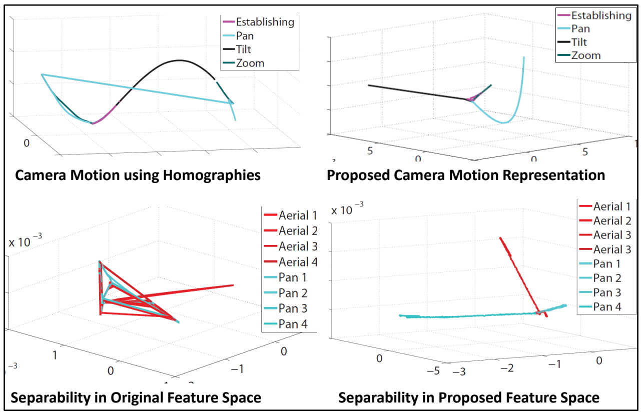
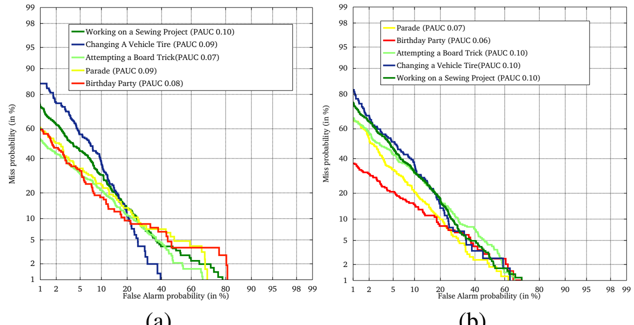


Figure 3: **Discriminativity of our proposed representation of shots in contrast to pure frame-to-frame homographies. Figures on top show four classes of shots (Establishing, pan, tilt and zoom) in both feature space. Bottom figures show clear separability of "aerial" and "pan" shot classes in the proposed feature space.**

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. 3). Multiple time series are then constructed from these mappings. 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 [3, 4], which demonstrates conclusive evidence towards its applicability in open-source video analysis (Fig. 4).

2.3 Intermediate representations

Designing intermediate representations on top raw features is very crucial for any recognition algorithm to handle outliers efficiently and reduce processing of large volumes of high dimensional data. We present an efficient alternative [1] to the traditional vocabulary based on BoVW methods used for visual classification tasks. Our representation (Fig. 5) 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. Since our proposed representation is based on MLE over a large set of supporting datapoints, we are able to capture more diversity in the data as opposed to conventional vector quantization based representations. This is indicated in Tab. 2.



(a) (b)

Figure 4: **Detection-Error Trade off (DET) curves for (a) 5 event classes best represented using our camera motion based features. (b) Curves for corresponding event classes obtained using Bag of SIFT features. .**

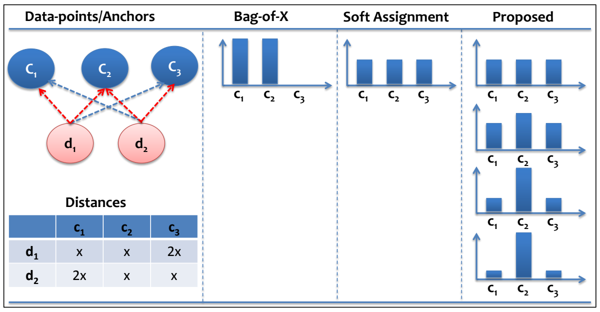


Figure 5: **Toy example contrasting the proposed representation against traditional BoVW and soft-assignment BoVW. Note that the proposed representation is initially identical to soft BoW but diverges since it maximizes an instance-level likelihood score.**

We integrate the above representation scheme to detect semantically accurate, human-understandable mid-level spatio-temporal concepts for modeling complex events. To this end, we introduce a benchmark dataset for spatio-temporal concepts, explicitly catering to the event recognition problem. This dataset consists of 62 mutually exclusive, concept categories over 10, 000 annotated audio visual samples extracted from NIST's TRECVID MED 2011 event corpus that replicates complex events observed in common video footages. Detectors are trained on the proposed representation [1] specific to each concept category on different information modalities (motion, static, and audio). This approach achieved respectable target detection [4] in the annual NIST TRECVID Multimedia Event Detection 2011 competition.

2.4 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 effectively exploited. To this end we propose the use of Linear Dynamical systems to create complex events. We hypothesize that a video depicting a complex event, can be expressed as an ordered vector time-series,

	Representation			
	Bag-of-Features		Anchors	
Modality	Avg. PAUC	AP(%)	Avg. PAUC	AP(%)
Static [SIFT]	0.2203	22.01	0.1949	23.52
Static [GIST]	0.2718	18.21	0.2223	18.15
Motion [Dollar]	0.1948	16.22	0.1735	19.24
Motion [STIP]	0.1869	17.31	0.1878	19.42
Motion [MBH]	0.1721	19.21	0.1639	20.21
Audio [MFCC]	0.3121	11.13	0.2936	11.12
Fusion [SIFT+MBH]	0.1615	19.41	0.1524	21.02

Table 2: **Spatio temporal concept detector performance evaluation summary: We summarize the performance of 62 concept detectors constructed from BoW (Baseline) and our proposed representation [1], across different feature modalities. Two different metrics – average area under DET curve (Avg. AUC), and average precision (AP) are used for evaluation. The lower the AUC measure (on a scale of $0 - 1$) the more reliable the detector while for AP, the greater indicates better performance.**

where each time-step is a vector containing confidences returned by a set of pre-trained spatio-temporal concept detectors [4]. Observing carefully, even for two visually similar events as shown in Fig. 6, we notice subtle differences in the concept evolution pattern for these two events – Repairing vs Cleaning an appliance.

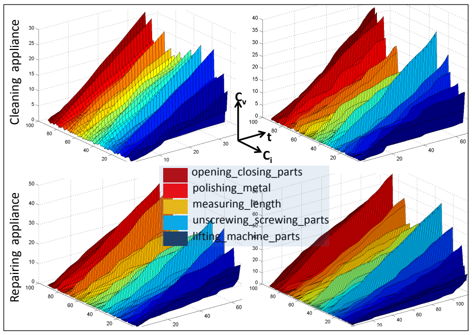


Figure 6: **Temporal dynamics of spatio-temporal concepts within a pair of complex events. Each column illustrates cumulative distribution of concept detector responses from a video (color coded to enhance readability). Top 5 relevant concepts in both events are indicated as inset legend.**

Hence, these evolving sequences can be modeled efficiently using linear dynamical systems (LDS). Now, direct estimation of LDS parameters – hidden state vector, state transition matrix etc. Conventional techniques use generative approaches – variants of discrete and continuous hidden markov models (HMM) to achieve this, however they are extremely sensitive to noise in training data. Since, concept detector responses are noisy, we propose to use alternative strategies that bypass this direct parameter estimation step. One technique is to construct a block-Hankel matrix for a vector time series which captures dependencies between each observation vector, within the context of the entire time-series. Eigenspaces of this matrix contains vital information which can be used to compute discriminative features to train any linear SVM classifier model, specific to a given event. Also, we can obtain meaningful statistics - such as periodicity, frequency, shift etc. and cluster the vector time-series into meaningful harmonic categories. Tab. 3 demonstrates the merits of our temporal models – Hankel Matrix based descriptors (HNK), Temporal Signatures (TS), and Late Fusion of HNK and TS based classifiers (FUSED), against some well known approaches – Discrete Cosine Transform based features computed from vector time-series (DCT), Discrete HMM (DHMM), and Continuous HMM (CHMM); from some preliminary experiments on TRECVID MED 2011 events collection dataset.

	Avg. Prec. from Method					
Event	**DCT**	**DHMM**	**CHMM**	**HNK**	**TS**	**FUSED**
E001	0.46	0.66	0.72	0.85	0.87	0.89
E002	0.44	0.64	0.71	0.89	0.91	0.91
E003	0.43	0.32	0.52	0.68	0.71	0.73
E004	0.39	0.39	0.39	0.61	0.59	0.59
E005	0.36	0.38	0.37	0.58	0.55	0.59
E006	0.34	0.38	0.51	0.87	0.87	0.85
E007	0.43	0.41	0.48	0.77	0.74	0.76
E008	0.67	0.69	0.71	0.88	0.89	0.87
E009	0.44	0.48	0.49	0.83	0.86	0.86
E010	0.38	0.48	0.51	0.74	0.75	0.75
E011	0.51	0.62	0.63	0.79	0.71	0.74
E012	0.37	0.73	0.68	0.78	0.76	0.78
E013	0.31	0.35	0.41	0.84	0.88	0.86
E014	0.34	0.46	0.48	0.68	0.68	0.67
E015	0.32	0.31	0.38	0.58	0.57	0.59
mAP	0.41	0.48	0.53	0.75	0.76	0.76

Table 3: **Average Precision scores (MED11EC) Performance of our proposed temporal features with contemporary methods that model temporal interactions on all 15 events in the MED11EC dataset.**

3. CONCLUSION

We presented a set of novel methodologies to perform semantic analysis of web videos. We introduced a principled decomposition of these videos into hierarchical components, highlighting our contributions in three key stages. As part of the first stage, we introduced two novel semi-global features which can be used to capture complementary information from videos. After that, we proposed an intermediate representation replacing the Bag-of-words model and demonstrated how this representation can be used to detect semantic concepts from videos. In the concluding section, we insinuated two discriminative feature spaces to model temporal interactions between detected concepts which can be efficiently integrated into existing classifiers for complex event recognition. In principle, the stages in the bottom-up approach suggested here, can be integrated in a common framework to perform semantic analysis of Internet videos.

4. REFERENCES

[1] S. Bhattacharya, R. Sukthankar, R. Jin, and M. Shah. A probabilistic representation for efficient large scale visual recognition tasks. In *IEEE CVPR*, pages 2593–2600, 2011.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE ICCV*, pages 1395–1402, 2005.

[3] H. Cheng, J. Liu, A. Hauptmann, S. Bhattacharya, M. Shah, and G. Friedland. SRI-Sarnoff AURORA System at TRECVID 2012: Multimedia Event Detection and Recounting. *NIST TRECVID and Workshop*, Dec. 2012.

[4] H. Cheng, H. S. Sawhney, A. Hauptmann, M. Shah, S. Bhattacharya, M. Witbrock, G. Friedland, R. Manmatha, and J. Allan. Team SRI-Sarnoff's AURORA System @ TRECVID 2011. *NIST TRECVID and Workshop*, Dec. 2011.

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

[6] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, November 2012.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *IEEE ICCV*, 2011.

[8] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE ICCV*, pages 432–439, 2003.

[9] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *IEEE ICPR*, pages 32–36, 2004.

[10] Y. Song, M. Zhao, J. Yagnik, and X. Wu. Taxonomic classification for web-based videos. In *IEEE CVPR*, pages 871–878, 2010.

[11] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. YouTubeCat: Learning to categorize wild web videos. In *IEEE CVPR*, pages 879–886, 2010.