21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Activity Detection in the Wild using Video Metadata

Scott McCloskey and Pedro Davalos
Honeywell ACS Labs, 1985 Douglas Drive North, Golden Valley, MN, 55422 USA
{*scott.mccloskey, pedro.davalos*}*@honeywell.com*

Abstract

We use video metadata to perform activity detection from videos in the wild, particularly the TRECVID dataset. Unlike previous activity datasets (KTH, Weizmann, UCF sports, etc.), this test set is assembled from videos captured with a wide range of cameras, resulting in videos with different frame rates, audio/video bitrates, and resolutions. Because these measures correlate with the quality of the camera, and because different camera hardware may be used to capture different events (e.g., people likely bring nicer cameras to weddings than on fishing trips), we expect that usable correlations exist between metadata and events. Using SVM-based classification of a feature vector of metadata features, we demonstrate that such correlations do *exist. While the performance of this method is worse than traditional visual features, we demonstrate that they compliment such approaches using score fusion.*

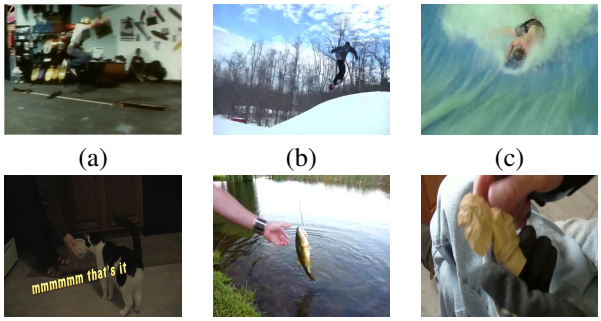

Figure 1. Example frames. Top row shows wide intra-class variation, with different clips labeled 'attempting a board trick': (a) skateboard, (b) snowboard (c) boogie board. Bottom row shows inter-class variation, with examples of 'feeding an animal', 'landing a fish', and 'woodworking'.

1. Introduction

Activity recognition is a topic of much recent interest in computer vision, and has a number of parallels with previous work in object or scene recognition. In general, both object and activity recognition use visual cues, from the low level (texture, motion, etc.) to high level semantics (parts models, people detection, etc.), which are extracted from the video/image pixels. While these cues are a natural way to address visual recognition, as they comport with our understanding of the human visual system, there are additional cues to visual recognition that compliment visual features.

In this paper, we describe the use of metadata for activity recognition of videos in the wild. The metadata describe the video - its resolution, length, etc. - in ways that one can use to infer what's depicted in therein. The length of a video clip, for instance, may be proportional to the complexity of the activity that it documents. Moreover, people may use a certain camera settings - or even a different camera entirely - when capturing video of a certain event. As an example, weddings are often filmed by professional videographers using relatively expensive equipment, whereas recreational events like fishing may be captured using lower quality devices such as cell phone cameras. We show that such relationships can help detect events in videos in the wild, and that such evidence is complementary to detection based on traditional visual features.

Our experiments are based on event detection on an archive of more than 300 hours (several thousand clips) of video collected and annotated for the Multimedia Event Detection (MED) task of TRECVID2011 [10][1]. The videos are uncontrolled with respect to camera motion, background clutter and human editing. As shown in Fig. 1, the event categories exhibit both wide intra-class variation (e.g. multiple semantic subclasses of *attempting a board trick*), broad inter-class variation, and

[1] http://www.nist.gov/itl/iad/mig/med11.cfm

rich temporal structure (e.g. *changing a vehicle tire*) which cannot be estimated from a single frame.

2. Related Work

The most related work to ours is the work by Boutell and Luo [1], who use EXIF metadata from still images to perform scene classification, particularly indoor-outdoor and sunset classification. Their work exploits relationships between the scene categories and metadata which are mechanical in nature, e.g. a camera's auto-exposure routine automatically selects a longer exposure time when capturing an image indoors, where lighting is generally limited. The EXIF metadata is generally discrete, e.g. 'flash fired' is a boolean, whereas our metadata are continuously-valued (see Sec. 4). The relationships we exploit are not artifacts of an automated system, as we are not aware of video cameras that select capture settings based on measurements of the scene. Instead, the human videographer sets our metadata, perhaps implicitly, when purchasing a particular device or manipulating its settings. In this sense, our metadata cues are likely weaker. Moreover, the event distinctions that we seek - say, differentiating between *feeding an animal* and *grooming an animal* - may be finer than the distinction between indoor and outdoor images.

As we will demonstrate, our approach is complimentary to activity recognition approaches based on the analysis of pixel intensities. There are many such approaches, which variously advance the representation of action features [13], the learning of relationships between features and activity categories [7], or the fusion of multiple features [5]. Activity recognition performance is often evaluated on the KTH [12] or Weizmann [4] datasets, which are ill-suited to our approach since all events are captured with the same camera and settings in order to control experimentation for the video's content. Likewise, the Hollywood [8] or UCF Sports [11] datasets are ill-suited for evaluation of our method, as they are trans-coded from broadcast video which is captured using all professional equipment. Instead, we evaluate our approach using the TRECVID data, which were collected from various online sites in the wild.

3. TRECVID and Metadata Collected

The TRECVID MED task was added to the annual evaluation in 2011 to assess the performance of event detection techniques on open source video clips. The evaluation provides training and testing video clips for several events. Our results are presented on the DEV-T data, containing five events:

- **E01** - Attempting a board trick
- **E02** - Feeding an animal
- **E03** - Landing a fish
- **E04** - Wedding ceremony
- **E05** - Working on a woodworking project

For each event type, approximately 100 positive training examples are given. The testing set consists of 4292 clips, comprising around 370 hours of video, with labeled instances of the events. All video is transcoded to MPEG4 video, and certain metadata (camera make, model, etc.) have been removed. As a result, the remaining metadata which we use in our experiments are: clip duration, video framerate, video bitrate, audio bitrate, and frame resolution.

4. Metadata-based Event Detection

Because the EXIF metadata used in [1] was discrete in nature, the use of a Bayes network was a natural way to perform classification. Even seemingly continuous values, such as exposure time, are quantized to have step sizes which are referred to by photographers as 'stops'. In our case, the metadata values are continuous, and range over a broad distribution. Framerate and resolution are the two partial exceptions to this. The distribution of frame rates is highly peaked around 30 FPS (NTSC video), 25 FPS (PAL), and 15 FPS (presumably sub-sampled NTSC), but other values are observed. Likewise, the distribution of resolution has pronounced peaks for standard video formats (VGA, PAL, SVGA, HD, etc.), and several non-standard values videos that were presumably edited post-capture.

In order to handle the continuously-valued metadata gathered from the TRECVID clips, we employ a Support Vector Machine (SVM) to perform activity detection. We use LibSVM [2], and train separate 1-versus-all classifiers for each of the five events. We add a bias term (a feature whose value is 1 for all clips) and use L1 regularization in learning. When training a classifier for event $N \in \{1, 2, 3, 4, 5\}$, we use the (roughly 100) training clips for that event as positive examples, and the training clips for events $\{1, 2, 3, 4, 5\} \backslash \{N\}$ as negative examples. Each video clip is represented by a feature vector containing the five metadata values, normalized to the range 0-1. Due to the low dimensionality of the feature vector, both training and prediction using metadata are quite fast. Moreover, since the feature vector can be constructed *without processing any of the pixels,* the overall metadata-based classification is extremely fast.

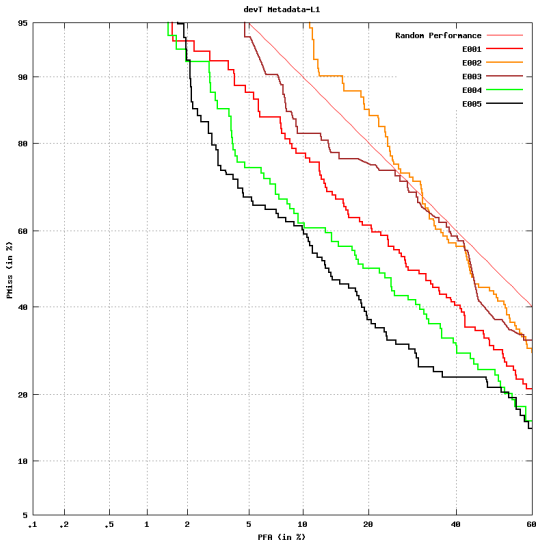

Figure 2. Metadata-based DET curves, using an SVM to detect 5 complex events in an archive of 4292 video clips.

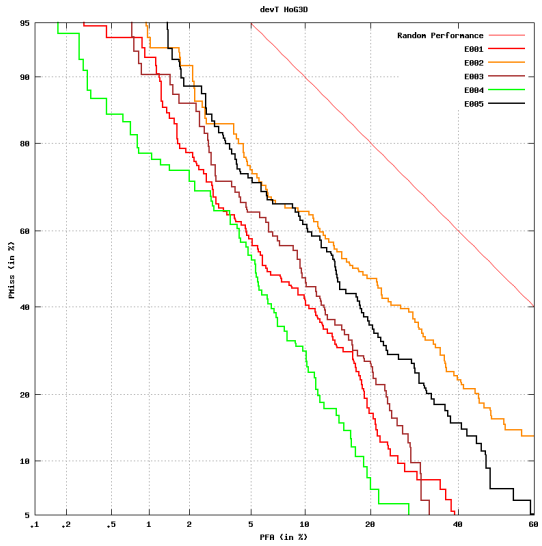

Figure 3. DET curves, using a BoW histogram based on HOG3D features.

Event	Metadata	HOG3D	Fused
Board trick	0.370	0.118	**0.113**
Feeding an animal	0.461	**0.262**	0.266
Landing a fish	0.418	0.132	**0.131**
Wedding ceremony	0.309	0.077	**0.076**
Woodworking	0.257	0.199	**0.168**

Table 1. Performance (AUC) on events, for metadata features, HOG3D, and fusion. Bold: best performance per event.

5. Experimental Results

Per TRECVID conventions, we present our experimental results with detection-error-trade-off (DET) curves, which are very similar to ROC curves but with nonlinear scaling on the two axes (false alarm probability, miss detection probability), so that the curves are more 'linear' [9]. Using the method described in the previous section, our detection performance is shown in Fig. 2. For comparison, Fig. 3 shows DET curves on the same events, using visual features. In particular, we compute densely-sampled Histograms of Oriented 3D Gradient (HOG3D) features [6] over the video, and quantize them to a Bag of Words (BoW) feature over a vocabulary size of 1000. Because DET curves may have several crossings, directly comparing the curves is tricky. In order to provide a summary metric, we compute the Area Under the Curve (AUC, where lower is better) on these DETs, which we present in Table 1.

From the curves and AUCs, we see that metadata-based classification has performance on the *woodworking* event which is comparable to the performance of the HOG3D feature. On three other events, metadata performance is worse than HOG3D performance, but is better than random chance. On the last event, *feeding an animal*, metadata performance is worse than both HOG3D features and comparable to random chance.

In addition to the differences in detection performance, it is important to understand the differences in computational complexity. Whereas the metadata features can be constructed on a standard PC in a negligable amount of time, extraction and quantization of HOG3D features is time-consuming (and proportional to the total video *duration* in the archive). The HOG3D features used here were computed at 1.43 seconds per video second (i.e., at 1.43x real time), meaning that feature computation took around 500 CPU-hours.

5.1. Fusion with Visual Features

Because of the low computational complexity of metadata-based event classification, it can be viewed as complimentary to traditional, pixel-based event detection. It is well known that committees of classifiers can be fused to provide improved performance, providing that the scores of the base classifiers are de-correlated. In order to test this, we use Maximum Figure of Merit (MFoM) fusion [3] to combine detection scores obtained from the metadata and HOG3D base classifiers.

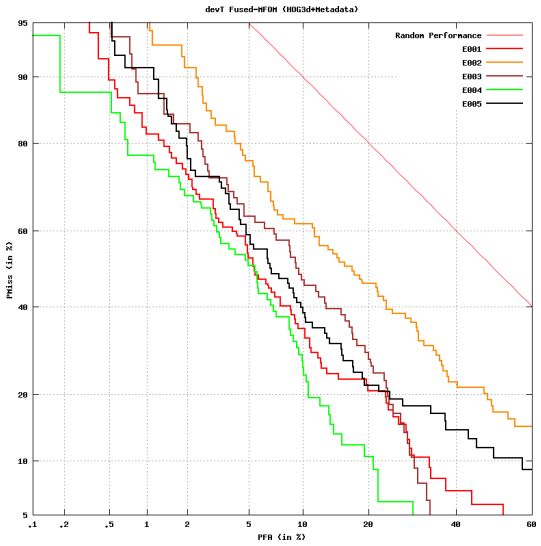

Figure 4. DET curves, using a fusion of metadata- and HOG3D-based detection.

Note that this requires that some of the scores used to produce Figs. 2 and 3 be used for fusion training; we use 20% for MFoM training, and the remaining 80% to evaluate performance. Fig. 4 and Table 5 (right column) show the results of this fusion, and we see that performance is significantly improved on *woodworking* while not being reduced on other events.

6. Analysis and Conclusions

We present an evaluation of metadata-based activity recognition from videos in the wild, inspired by previous work showing scene recognition from EXIF camera metadata. Whereas that study used a rich set of metadata features with correlations introduced by the mechanical design of a camera, our features are relatively poor since camera metadata has been removed. However, we show that the remaining metadata can be used to detect complex activities in a large video archive. While performance is generally worse than methods using traditional visual features in a bag of words representation, we demonstrate that metadata-based analysis compliments this approach. Should an activity recognition test set become available with additional metadata features, we would expect improved performance from such a method.

Acknowledgements

The authors would like to thank Greg Mori and Arash Vahdat for providing the HOG3D features.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

References

[1] M. Boutell and J. Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *Computer Vision and Pattern Recognition*, 2004.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[3] S. Gao, C. Lee, and J. Lim. An ensemble classifier learning approach to roc optimization. In *Int'l Conf. on Pattern Recognition*, 2006.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[5] M. I. Jordan. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[6] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008.

[7] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011.

[8] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition*, 2009.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *European Conf. on Speech Communication and Technology*, 1997.

[10] P. Over, G. Awad, J. Fiscus, B. Antonishek, A. F. Smeaton, W. Kraaij, and G. Quenot. TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2010*. NIST, USA, 2011.

[11] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition*, 2008.

[12] C. Schldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Int'l Conf. on Pattern Recognition*, pages 32–36, 2004.

[13] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *Computer Vision and Pattern Recognition*, 2011.