Internet Video Category Recognition

Grant Schindler
Georgia Institute of Technology
schindler@cc.gatech.edu

Larry Zitnick
Microsoft Research
larryz@microsoft.com

Matthew Brown
University of British Columbia
mbrown@cs.ubc.ca

Abstract

In this paper, we examine the problem of internet video categorization. Specifically, we explore the representation of a video as a "bag of words" using various combinations of spatial and temporal descriptors. The descriptors incorporate both spatial and temporal gradients as well as optical flow information. We achieve state-of-the-art results on a standard human activity recognition database and demonstrate promising category recognition performance on two new databases of approximately 1000 and 1500 online user-submitted videos, which we will be making available to the community.

1. Introduction

In this paper, we examine the general problem of internet video categorization. We make no assumptions about the videos we attempt to categorize: each video may be recorded from a hand-held video camera, a cellphone, a webcam, a television broadcast, or even an animated cartoon. An excellent source of such a wide variety of videos is the growing number of user-submitted video websites which have become popular over the last few years. As such, we explore the video categorization problem on two new databases of approximately 1000 and 1500 online user-submitted videos which we will be making available to the community.

Specifically, we explore the representation of a video as a "bag of words", based on recent successes of this approach in both activity recognition [2, 13, 17] and video scene retrieval [15]. In addressing the general problem of video category recognition, we open up the possibility of directly comparing the motion-based methods of activity recognition with the appearance-based methods used in object, scene, and location recognition tasks [3, 10, 9]. One of the goals of this work is to determine how general video category recognition differs from the more commonly studied activity recognition problem. For this reason, we make as few assumptions as possible about what types of interest points and descriptors are well suited for the task of video category recognition. Instead, we define a number of spatial, temporal, and spatio-temporal interest points to be combined with descriptors based on gradient orientation and optical flow. The performance of each combination is then evaluated for the video recognition task.

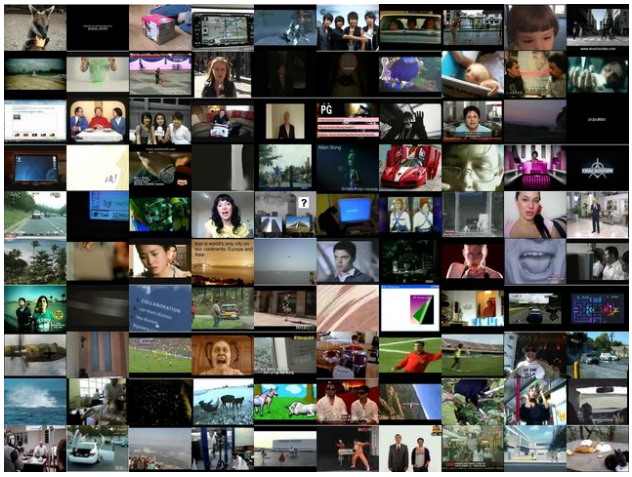

Figure 1. 100 videos from a database of approximately 1500 user-submitted online videos across 15 categories – we perform video category recognition in a bag of words framework.

The systematic enumeration and evaluation of video interest point and descriptor combinations is a primary contribution of our work. This approach is justified by a new state-of-the-art performance result on the KTH human activity recognition database, and promising video category recognition performance for two user-submitted video databases. In addition we present a novel temporally-varying optical flow histogram descriptor for both space-time and temporal interest points, and a novel temporal interest point detector based on optical flow magnitude.

1.1. Related Work

Inspired by the success of local features in static 2D images [11], much recent work in activity recognition has been based upon the use of local features detected in the 3D space-time volume defined by a video. Space-time interest points were first introduced in [8] by extending Harris cor-

ner detection to find large changes in image intensity in both space and time. Another space-time interest point detector based on 1-D Gabor filters was introduced in [2].

Once space-time interest points are detected, descriptors of the space-time volume around the interest point have traditionally been based on either image gradients or optical flow. Oriented histograms of differential optical flow were used in [1] to detect humans in video. Simple volumetric features based on differences of optical flow in neighboring regions were used for visual event detection in [6]. In [7] spatio-temporal shape and flow correlation is performed on over-segmented videos to recognize human activities.

Our approach shares much in common with the work of Dollar et al. [2]. In contrast to previous approaches to behavior recognition which relied upon detection and segmentation, [2] used sparse spatio-temporal features for behavior recognition, and [13] extended the same methods to the unsupervised case. The work in [2] explored a number of descriptors for spatio-temporal interest points, including those based on space-time gradients, optical flow, and image intensities. These were placed into a bag-of-words framework where descriptors are clustered to form a vocabulary and recognition is performed in the space of word-frequency histograms. Here, we adopt the space-time interest point detector and space-time gradient PCA descriptor of [2].

In addition to the above activity recognition work, there is also a body of work on content-based video retrieval as exemplified by the TREC Video Retrieval Evaluation project [16] which focuses on the related problem of segmenting and detecting shots of specific objects and settings in multi-hour videos. The Video Google work of Sivic and Zisserman [15] is an example of such an approach, and in this work, we compare activity-recognition-inspired approaches against purely appearance based approaches such as this.

2. Approach

Given a large database of videos (each labeled with a category) and a new unlabeled video, we would like to infer the category of the new video. For this task, we have chosen to convert each video to a *bag-of-words* representation. To accomplish this, we detect interest points in each video and compute a descriptor for each interest point. We use k-means to cluster the descriptors from all videos into a vocabulary consisting of N words. For each video we construct a normalized word-frequency histogram that describes how often each vocabulary word occurs in the video. This histogram can be viewed as an N-dimensional representation of the entire video and all classification is performed in this space. In the next section, we discuss the specific interest points and descriptors used to form a variety of these vocabularies, and therefore a variety of *bag-of-words* representations for each video.

3. Interest Points

For any given video, we can detect three kinds of interest points: spatio-temporal, temporal, and spatial. While spatial interest points (e.g. those used with SIFT [11]) and spatio-temporal interest points (as used in activity recognition) enjoy widespread use, purely temporal interest points have not been popular in video analysis (though [17] uses them as an intermediate stage in spatio-temporal interest point detection). We include temporal interest points here for completeness of the evaluation. Each type of interest point is explained in detail below.

3.1. Spatio-Temporal Interest Points

Spatio-temporal interest points are detected using the 1-D Gabor filter method of [2]. They are detected at a single scale, using $\sigma = 2.0$ pixels. The Gabor filter responses form a 3D volume in space and time and we take the maxima in this space as interest point candidates. We reject any interest points whose filter responses are less than 1.1 times greater than the filter responses of all neighboring points. The resulting interest region is a volume measuring 48x48 pixels in spatial extent and 30 frames in time, centered on the interest point. A typical 400x300 pixel, 500-frame video generates 3000 spatio-temporal interest points.

3.2. Temporal Interest Points

Temporal interest points are detected at multiple temporal scales and always include the entire spatial extent of all frames that fall within the corresponding region of temporal interest. We first compute the optical flow at every frame of the video using the Lucas-Kanade [12] method on 3x3 pixel regions spaced at 4 pixel intervals in both the x and y directions. We define the motion m_t in a given frame t as the sum of the magnitude of the optical flow at all pixels in the frame. We construct the 2D difference of Gaussian (DoG) [11] based on the 1D motion signal $m_{1..T}$ and search for extrema in this space. We examine 20 levels in the DoG. At level s, the motion signal is convolved with a zero-mean Gaussian with $\sigma = 1.2^s$, where the number of frames included in the temporal region of interest is 8σ. A typical 400x300 pixel, 500-frame video generates roughly 300 temporal interest points, mostly at smaller scales.

3.3. Spatial Interest Points

Spatial interest points are computed for individual frames of video using the familiar Difference of Gaussian (DoG) method in [11]. They are computed at multiple spatial scales, but do not have any extent in time and exist only at a single frame. In all results reported below, spatial interest points are detected for one in every 15 frames of video. A typical 400x300 pixel, 500-frame video generates 6000 spatial interest points.

4. Descriptors

The interest points defined above define space-time volumes (as described in section 3) to which we may assign a number of descriptors. Note that for space-time interest points, the space-time *volume* of interest is explicitly defined above. For temporal interest points, the space-time volume of interest encompasses the entirety of the video frame for every frame in the temporal region of interest. For spatial interest points, the space-time volume of interest includes only the spatial region immediately surrounding the interest point (as defined by the scale of the point) and only in the frame in which the point was detected.

4.1. Space-Time Gradient

At every pixel in the space-time volume surrounding an interest point, we compute the image gradient in the x, y, and time directions. While the x and y gradients are typical image edges, the time gradients describe the change in a pixel's intensity from one frame to the next. If we concatenate all x, y, and time gradients together the result is a descriptor with thousands of dimensions. In order to reduce the dimensionality of the descriptor and to make it robust to small variations, we adopt two methods: principal components analysis (PCA) and histogramming.

4.1.1 Space-Time Gradient PCA

As in [2], we perform PCA on the large concatenated vector of x, y, and time gradients to get a small number of vectors that describe the main modes of variation of the data. We take the first 50 principal components and describe the region by the amount to which the gradients project onto each of these principal components. For computational efficiency, we downsample the space-time volume of interest to 5x5 pixels by 11 frames and so have to compute only 4x4x10=160 pixel differences, multiplied by 3 to account for each of the x, y, and time gradients. The descriptor is 50 dimensions and normalized to unit length.

4.1.2 Space-Time Gradient Histograms

Alternatively, we can compute histograms that describe the orientation of edges in the xy, xt, and yt planes that cut through the space-time volume of interest. We can understand these histograms by analogy with the SIFT descriptor, which captures the distribution of edge orientations in an image region by computing a direction and magnitude corresponding to the x and y gradient at each pixel in the region. Because we deal with a space-time volume, we can histogram along the yt and xt planes in the exact same way, resulting in 3 histograms of 10 bins each, resulting in a 30-dimensional descriptor. Each of the 3 histograms is normalized to unit length independently.

4.2. Optical Flow Orientation Histogram

Here we use the same optical flow computation that was used in detecting temporal interest points. For a given space-time volume, we uniformly divide the volume into 8 temporal regions. For each region, we compute an 8-bin histogram of optical flow orientation. This results in an 8x8=64 dimensional descriptor of the motion in the space-time volume of interest, which we normalize to unit length.

4.3. SIFT

The descriptor for all purely spatial interest points is the standard SIFT descriptor from [11]. Video searching via spatial interest points with SIFT descriptors was a method originally introduced in [15].

5. Classification

We use two classification techniques: 1-nearest neighbor (NN) and centroid-based document classification [4] which performs extremely well in textual document classification by simply representing a class with the centroid of all documents that belong to the class. In all cases, we use the χ^2 distance metric [2]:

$$Dist(a,b) = \sum_{i=1}^{N} \frac{(a_i - b_i)(a_i - b_i)}{2(a_i + b_i)}$$

to compare the N-dimensional normalized word-frequency histograms which represent any two videos a and b. In the NN case, we compute the distance from a query video to all database videos and assign to the query video the category of the minimum-distance database video. In the centroid case, we represent a category as the mean of all histograms of videos in the given category. For M categories, we then have just M histograms to compare with the histogram of the query video. These two techniques were chosen for their speed in both training and classification.

6. Results

We tested several combinations of interest points and descriptors on three separate classification tasks. We combine spatio-temporal interest points with the PCA and histogram versions of the space-time gradient descriptor as well as the optical flow orientation histogram. We combine temporal interest points with space-time gradient PCA and the optical flow orientation histogram. Since the spatial interest points have no temporal extent, we only combine them with the SIFT descriptor.

6.1. Data Sets

We test the above methods on three data sets. The first set consists of 1483 videos assembled by downloading the

Interest Point	Descriptor	KTH	Sports	General
ST	STG-PCA	**82.6**	27.2	**21.8**
ST	Flow	80.6	28.7	19.5
ST	STG-Hist	73.9	25.7	19.0
T	STG-PCA	60.2	26.3	17.3
T	Flow	52.1	25.2	18.6
S	SIFT	43.8	32.0	17.8

Table 1. Centroid-Based Classification Results. In all tables, we use S for spatial, T for temporal, ST for spatio-temporal, and STG for space-time gradient. In all tables, the amounts reported are the percentage of the data correctly classified. The best individual-vocabulary results for the KTH and General video data sets reside here, and are indicated in bold text.

Interest Point	Descriptor	KTH	Sports	General
ST	STG-PCA	71.2	35.9	20.6
ST	Flow	78.6	31.3	18.2
ST	STG-Hist	70.6	37.2	21.8
T	STG-PCA	45.1	28.3	15.0
T	Flow	46.2	30.4	16.8
S	SIFT	42.5	**38.3**	18.9

Table 2. Nearest Neighbor Classification Results. The best individual-vocabulary results for Sports reside here, indicated in bold text. See Table 1 for an explanation of the notation used in this table.

most popular 100 videos from each of the 15 general video categories on an online video sharing website. The categories are *animals, animation, autos, blogs, comedy, commercials, entertainment, games, movies, music, people, politics, sports, technology,* and *travel*. We call this the *General* video data set.

The second set of 967 videos consists of the top 100 videos returned from each of 10 searches for the names of different sports. The sports are *baseball, basketball, bowling, football, karate, skiing, soccer, surfing, swimming,* and *tennis*. We call this the *Sports* video data set.

The videos in these two datasets are all 400x300 pixels and we perform classification based on the first 500 frames of each video (i.e. a 60 megapixel space-time volume). All videos were categorized, sometimes "incorrectly", by the users who submitted them – thus a number of potentially irrelevant videos for each of the 15 general categories and 10 sports are included in both the training and testing sets of all results presented below.

The third data set is the KTH Human Motion Database, consisting of 598 videos showing 25 participants performing 6 actions in 4 different situations. The six actions are *walking, jogging, running, handclapping, handwaving,* and *boxing.* We include this standard dataset so that we may compare our method directly against other similar approaches extracting descriptors from video for the purpose of classification.

6.2. Experiments

We conducted 78 experiments testing various combinations of interest points and descriptors. In all experiments, we use a 600 word vocabulary. We tested vocabularies of size 300, 500, 600, 700, and 800, and achieved the best performance with 600. In addition, we test concatenations of the word-frequency histograms (independently normalized) of several descriptors. For example, we append the 600-dimensional histogram of space-time gradient PCA word frequencies with the 600-dimensional histogram of optical flow orientation histogram word frequencies to arrive at a new 1200-dimensional representation of the video by simple concatenation. In our experiments, we find that this alone can increase the categorization performance by a significant amount. For numerical results, see Tables 1-4. In the next section, we provide more details and discussion of the experimental results.

7. Discussion

The relatively similar performance of the space-time gradient PCA and optical flow histogram descriptors was unexpected given the findings in [2] that space-time gradient PCA significantly outperforms optical flow histograms in activity recognition performance. The primary distinction between our flow histograms and those used in previous work is that we compute the optical flow histogram for multiple temporal regions within a space-time volume rather than for multiple spatial regions – thus our descriptor emphasizes how motion changes over time. Note that in several instances in the tables above, our optical flow histogram approach out-performed space-time gradient PCA as a descriptor for space-time interest points.

The results of combining space-time gradient PCA and optical flow are also surprising and significant. In previous

work [2], the superiority of space-time gradient PCA was used as justification to throw out optical flow information altogether. However, our exhaustive experimentation shows that concatenating the bags of words for these two descriptors results in significant improvement over either descriptor individually. Note that it is not the case that concatenation of any two descriptors will necessarily lead to better classification rates, as in several cases above combining SIFT with other descriptors leads to a net decrease in performance.

Another surprising result is that for the Sports dataset, the best individual performance was achieved by a histogram of SIFT features detected in the individual frames of the video – that is, by ignoring all temporal information and treating the video as a group of unordered images. It is somewhat unintuitive that the motion information in action-packed sports videos is less useful than appearance for video categorization. However, we can make sense of this result by observing that videos within a single sports category have more uniform visual appearance than general videos, due to the highly structured physical environments in which most sports take place. In contrast, SIFT performs poorly on the KTH activity recognition database in which a person's appearance is unrelated to the activity they are performing.

In contrast to the Sports results, on both the KTH and General datasets, the best individual performance is achieved using the space-time interest points with space-time gradient PCA descriptor. In a broad sense, the KTH result confirms the findings of [2], while the General result provides evidence that the success of space-time interest points and space-time gradient PCA descriptors carries over to the case of general videos, and is not restricted to the highly structured environments and shooting conditions of most human activity recognition work. The success of this particular combination of interest point and descriptor for general videos is significant as a large motivator of this work was to discover whether activity recognition results would carry over to online video categorization.

In general, the purely temporal interest points performed more poorly than spatio-temporal interest points, though purely temporal interest points do out-perform purely spatial interest points (SIFT) for General videos under centroid-based document classification.

Finally, note that with respect to descriptors for spatio-temporal interest points, all three types of descriptor (optical flow orientation histogram, space-time gradient PCA, and space-time gradient histogram) outperformed the others on at least one dataset using one classifier. Thus, at the end of this experiment, we cannot say that one of these descriptors should be used universally to the exclusion of the others. This leads us to our second set of experiments in which we combine multiple interest-point-descriptor pairs by simply concatenating their normalized histograms.

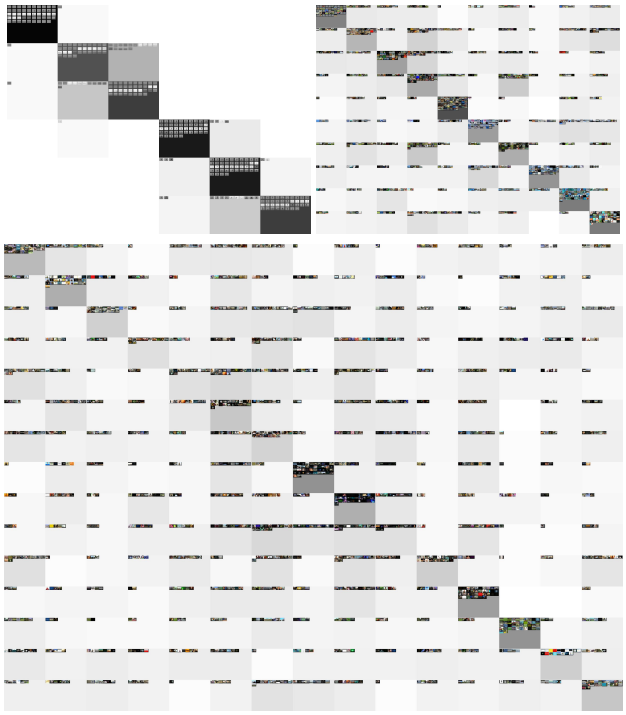

Figure 2. Confusion Matrices for KTH, Sports, and General video data sets (top left, top right, bottom). Each cell indicates which videos from one category (the row) were classified as belonging to another category (the column). All categories are in the order indicated in Section 6.1. Best viewed at high magnification.

7.1. Combinations of Features

For the KTH, Sports, and General datasets, recognition performance was raised by 4.7%, 4.4%, and 5.1% respectively by concatenating the normalized histograms of two or more interest-point-descriptor pairs (see Tables 3 & 4). This is quite a significant result, and as discussed above, this improvement is not the inevitable result of such a concatenation. Instead, it reflects the fact that we have identified certain descriptors which produce mutually beneficial encodings of the same data (i.e. that each descriptor alone may result in a different subset of the dataset being correctly classified.) The consistency of this result across all three video databases suggests an extremely simple method of uniting a wide variety of interest points and descriptors into a common framework while boosting performance at the same time.

7.2. KTH Activity Recognition Database

Using the above approach, we have exceeded the current state of the art results for the KTH database by achieving a categorization rate of 87.3%, compared to 81.2% in [2], 81.5% in [13], and 86.83% in [14] under exactly the same experimental conditions. We note that [5] recently achieved a rate of 91.7% under different experimental conditions.

STG-PCA	Flow	STG-Hist	SIFT	KTH	Sports	General
x			x	79.9	32.3	24.8
	x		x	82.3	34.7	24.1
		x	x	75.9	30.5	22.0
	x	x	x	78.6	31.5	25.2
	x	x		78.6	27.7	22.5
x	x			**87.3**	31.5	25.3
x	x		x	86.9	35.1	**26.9**

Table 3. Centroid-Based Results for Concatenated Word-Frequency Histograms. In this table, all descriptors are combined with spatio-temporal interest points, except for SIFT which uses spatial interest points. Each row indicates a different combination of vocabularies in the concatenated word-frequency histogram. An "x" under a descriptor's name indicates that a vocabulary based on this descriptor was included in the test for the current row. A combination of SIFT, optical flow, and space-time gradient PCA vocabularies perform best for the Sports and General data sets, while KTH performance is actually harmed by the inclusion of SIFT features.

STG-PCA	Flow	STG-Hist	SIFT	KTH	Sports	General
x			x	71.6	41.5	22.3
	x		x	74.9	39.7	23.1
		x	x	72.6	42.2	23.0
	x	x	x	75.9	41.1	24.3
	x	x		77.2	37.6	22.6
x	x			82.9	38.2	23.7
x	x		x	82.6	**42.7**	24.3

Table 4. Nearest Neighbor Results for Concatenated Word-Frequency Histograms. See Table 3 for an explanation of notation.

Specifically, we use the half of the database corresponding to the *d1* and *d3* videos for 25 subjects performing 6 actions each. When classifying a given test video from one subject, only videos from the 24 other subjects are used to train the classifier. In all tests, the query was excluded from unsupervised vocabulary building. The principal component vectors used in the space-time gradient PCA descriptor (for all videos in the KTH, Sports, and General datasets) were derived from 6 videos in the unused half of the KTH dataset.

8. Conclusion

In this paper, we have described three main contributions. We exceed the state of the art on the standard KTH human activity recognition database. We explore a wide range of combinations of spatial, temporal, and spatio-temporal interest points with a number of descriptors based on both motion and appearance and show that a combination of motion and appearance descriptors outperforms either individually on several datasets. Finally, we are making available a unique dataset of online videos.

References

[1] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[3] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[4] Eui-Hong Han and George Karypis. Centroid-based document classification: Analysis and experimental results. In *Euro. Conf. on Principles of Data Mining and Knowledge Discovery*, pages 424–431, 2000.

[5] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[6] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[7] Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Visual Surveillance Workshop*, 2007.

[8] Ivan Laptev and Tony Lindeberg. Space-time interest points. *ICCV*, page 432, 2003.

[9] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

Figure 3. Nearest Neighbor Matching Results. For the query videos on the left, the nearest neighbor match using spatio-temporal interest points with a combination of STG-PCA and Flow descriptors is shown on the right. The corresponding word frequency histograms for the left and right videos are shown on the top and bottom, respectively, just below each matching pair. Each video is visualized here by a single frame taken half-way through the video.

Figure 4. Nearest-Neighbor Categorization Failures. Some categorization failures are clearly caused by visually similar videos being placed into multiple categories for semantic reasons, while other failure cases are not so easily explained away.

[10] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[11] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[12] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[13] J.C. Niebles, H. Wang, H. Wang, and L. Fei Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, page 1249, 2006.

[14] S. Savarese, A. Del Pozo, J.C. Niebles, and L. Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, 2008.

[15] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[16] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *International Workshop on Multimedia Info. Retrieval*, 2006.

[17] S. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. *ICCV*, 2007.