Features and Automatic Search at The University of Iowa

David Eichmann and Dong-Jun Park

Institute for Clinical and Translational Science
The University of Iowa
david-eichmann@uiowa.edu

Abstract

High Level Feature Extraction

Approach – SVM fusion of color moment, edge direction and wavelets:

- UIowa07Feat1 – color moment
- UIowa07Feat2 – edge direction
- UIowa07Feat3 – wavelet, level 1
- UIowa07Feat4 – wavelet, level 2
- UIowa07Feat5 – wavelet, level 3

No distinction in performance for task as defined, although performance is superior to our runs for TRECVID2006.

(Automatic) Search

Approach – fully automatic search exploring the impact of using frames sampled at 2 second intervals rather than keyframes:

- UIowa07AS01 – required ASR using noun phrases and entities
- UIowa07AS02 – required non-ASR run - color histogram * edge * distance
- UIowa07AS03 – color histogram similarity
- UIowa07AS04 – product of maximal histogram, edge, distance similarities for shot samples (one every 2 seconds)
- UIowa07AS05 – product of maximal histogram, edge, distance similarities for shot samples (one every 2 seconds) boosted by NPs and entities

Our best runs were for simple color histogram similarity, although the sampled interval product was very close in performance. Our worst performing run was the text only, indicating the impact that ASR and MT can have on a non-news corpus.

1 – High-Level Feature Extraction

Our work in feature extraction this year comprised two phases, generation of a set of low-level features and fusion of these features using support vector machines.

Low-Level Feature Extraction

Color Moment

We adopted the approach of Stricker and Orengo, where the color composition of an image can be viewed as a color distribution that can be described by its moments [4]. They proposed to use three low-order moments. The first order moment is the mean value of a color channel. The second and the third order moments are the standard deviation and the skewness of each color channel. We used LUV color space and generated the color moments on the whole image. This resulted in 9 bins: a bin for each of three color moments for each color space.

Edge direction

Edge is frequently used feature for automatic visual data tagging. Since the global color histogram does not provide any spatial or shape information, there have been intense research going on to utilize effective shape-describing

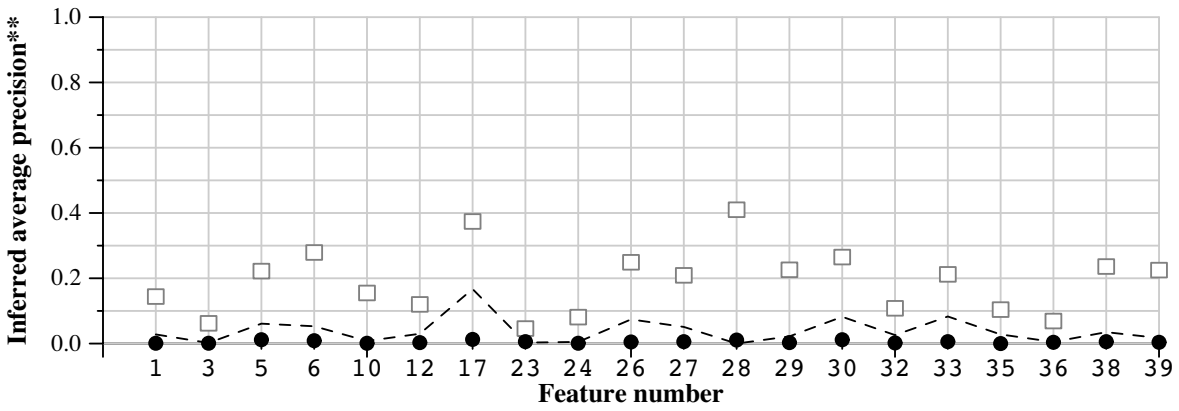


Figure 1: UIowa07Feat1

feature in visual data. One way is to extract edge by detecting sharp changes in the luminous intensity. Generally, this can produce two sets of information: edge magnitude and orientation.

To detect the edge orientation, a frame is converted to grayscale. In the next step, the image is convolved with a set of convolution kernels, each of which is sensitive to edges in a different orientation. For each pixel in the image, the local edge gradient magnitude is estimated with the maximum response of all eight kernels at this pixel location [1]. This covolution kernel is constructed so that each kernel is sensitive to certain orientation. We used a set of eight Sobel kernels to detect eight directions. The number of bins is 9: 8 for each of eight orientations (0, 45, 90, 135, 180, 225, 270 and 315 degrees) and one bin for non-edge points. The resulting histogram is normalized by the number of all pixels.

Wavelet

Wavelets are a mathematical tool for hierarchically decomposing functions or signals and have been applied to visual data analysis and compression since early 1990s. For given signal (or pixel value, in our case), a filter bank is applied by averaging and differencing coefficients, which filters out either high detail or low detail portions of the data. This allows us to determine the levels of detail to be present in the visual data [2]. For two-dimensional image data, a one-dimensional decomposition is performed on each row of the image, followed by a one-dimensional decomposition on each column of the result. This produces four frequency sub-bands of LL, LH, HL and HH. Here, L denotes low frequency, and H denotes high frequency. Thus, HL denotes high frequency in x-direction (row) and low frequency in y-direction (column).

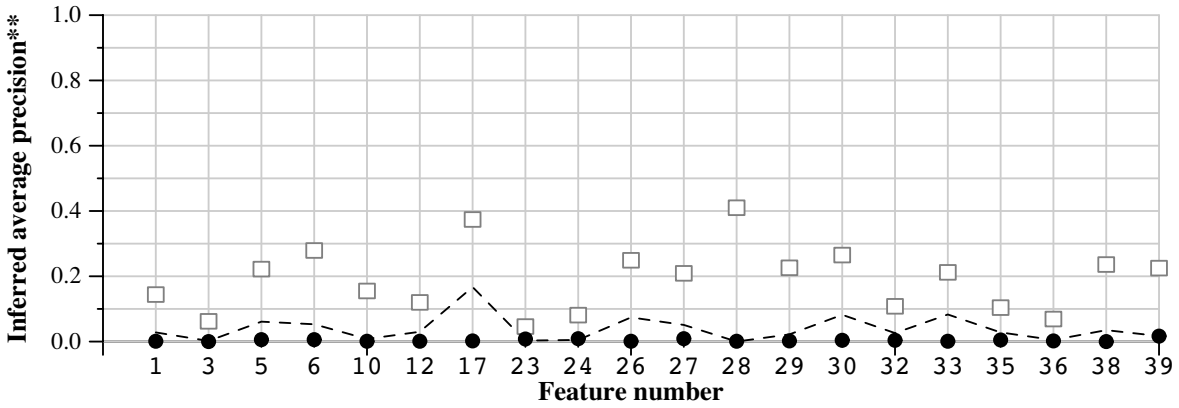
We performed 3 levels of a wavelet decomposition for each frame and calculated the energy level for each scale, which resulted in 10 bin feature data.

Support vector machine

For this year's TRECVID evaluation, we opted for using global visual features only. The entire feature set produced a feature vector of only 28 dimensions. We trained SVM classifier for each high-level concept with this feature vector and predicted the label of each shot in the test set with the trained SVM classifier. Figures 1-5 show the results of our official runs.

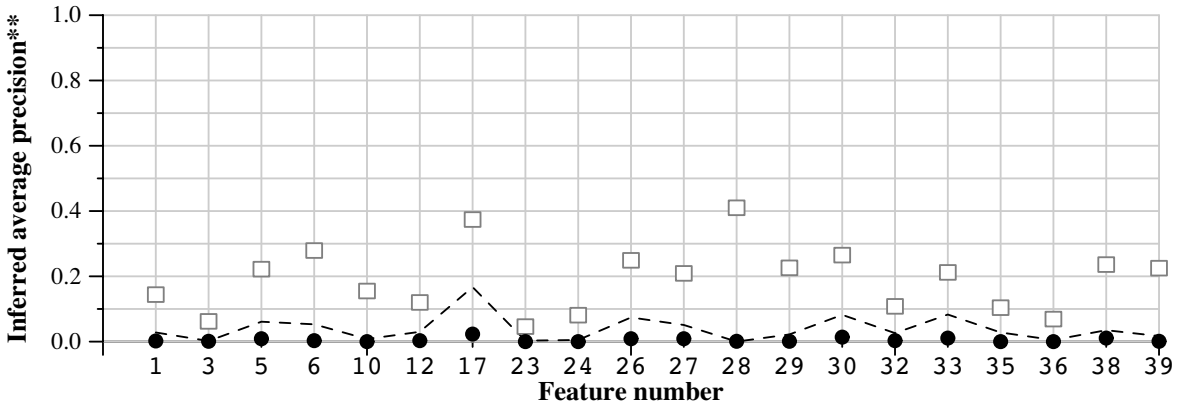
2 – Automatic Search

Our goal for this year's automatic search task was to explore the impact of keyframe selection on system performance, with a secondary aim of assessing the added value of machine-translated ASR, given the potential for high noise levels in this signal when compared to the scripted broadcast speech from recent TRECVID data. We generated two baseline runs, one solely on the MT ASR (submitted for condition 1) and one solely on the provided keyframes, using a similarity measure based upon a histogram/edge/distance product (and not submitted for official evaluation).



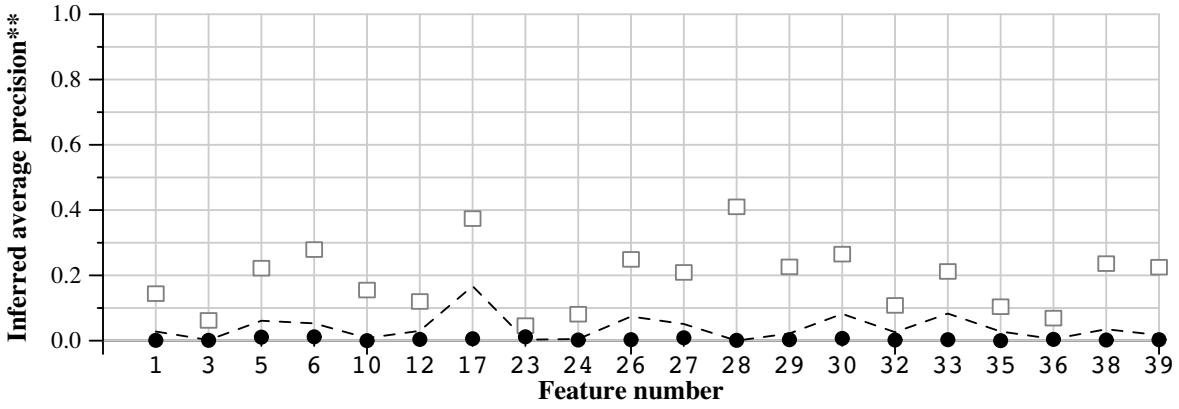
Run score (dot) versus median (---) versus best (box) by feature

Figure 2: UIowa07Feat2



Run score (dot) versus median (---) versus best (box) by feature

Figure 3: UIowa07Feat3



Run score (dot) versus median (---) versus best (box) by feature

Figure 4: UIowa07Feat4

We then generated frames at uniform, 2-second intervals for the collection and proceeded to treat them as fine-grained keyframes. Our rationale here was that given the long-duration and 'less-to-the-point' nature of this year's collection (when compared to news footage), there was significant likelihood that relevant activity could occur in a shot and not have it appear in the small number of keyframes officially generated for the collection.

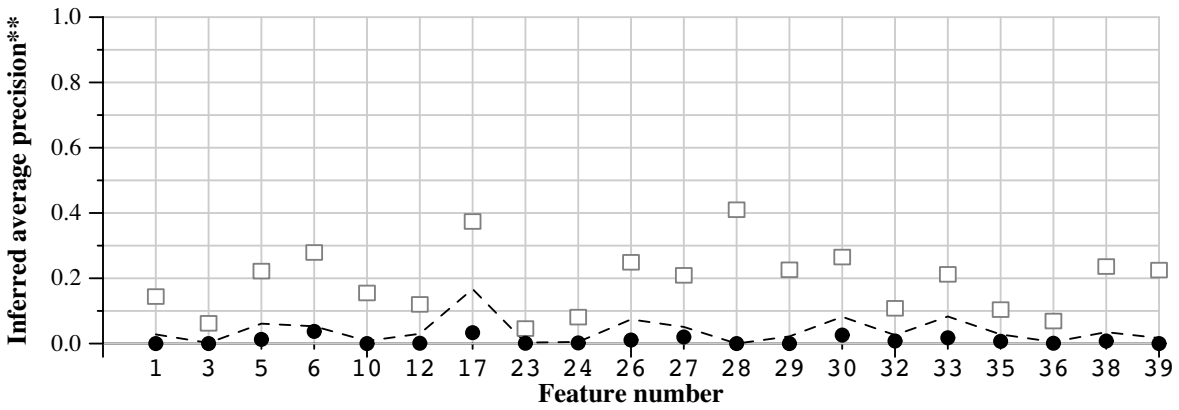


Figure 5: UIowa07Feat5

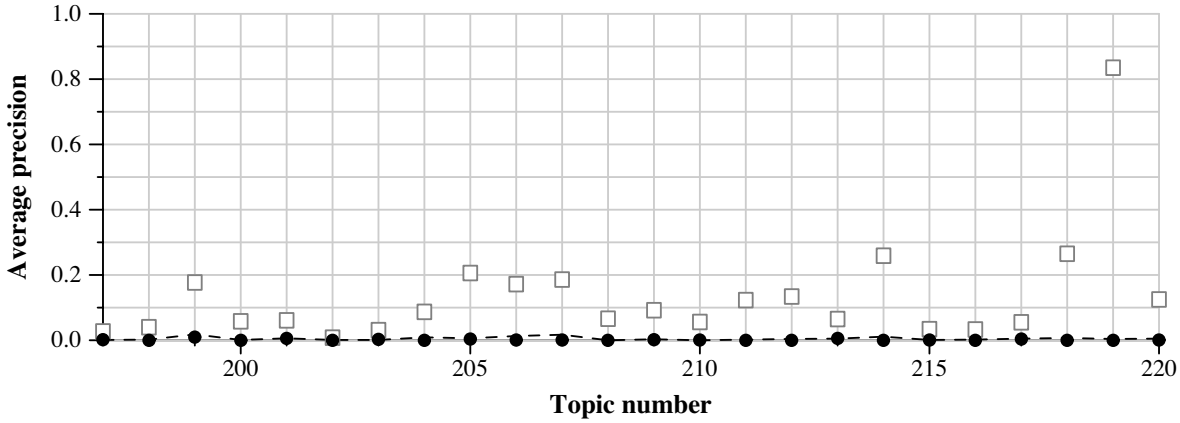


Figure 6: UIowa07AS01

A set of runs against the sampled-frame data were then generated using the same product measure as the baseline, one using the maximal match for each product term for each shot, one involving a simple color histogram comparison and one using the baseline product boosted with the baseline ASR scores. The talk will present an analysis of relative performance and a rational for the superior performance of the histogram configuration, closely followed by the maximal term product configuration.

As can be seen by Table 1 and Figures 6-10, overall performance is fairly modest.

Table 1: Search Results

Run	Description	Mean P @ T	Mean AP
UIowa07AS1	required ASR w/ noun phrases and entities	0.012	0.002
UIowa07AS2	required non-ASR run - color histogram * edge * distance	0.018	0.003
UIowa07AS3	color histogram similarity	0.024	0.005
UIowa07AS4	product of maximal histogram, edge, distance similarities for shot samples (one every 2 seconds)	0.022	0.003
UIowa07AS5	product of maximal histogram, edge, distance similarities for shot samples (one every 2 seconds) boosted by NPs and entities	0.019	0.005

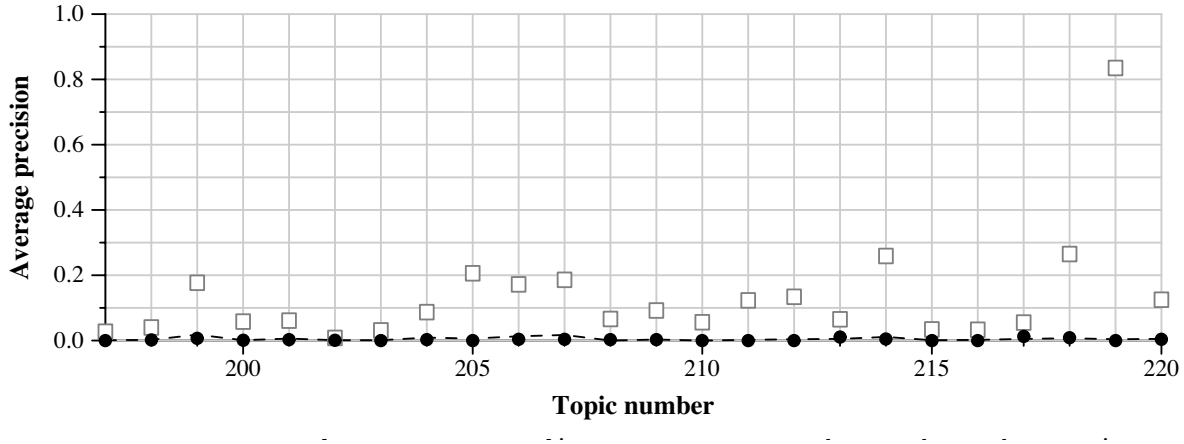


Figure 7: UIowa07AS02

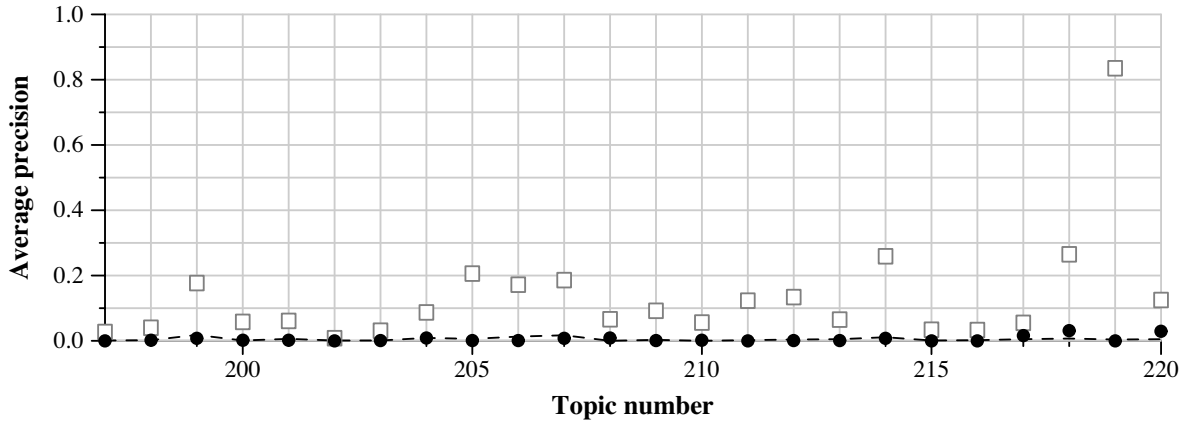


Figure 8: UIowa07AS03

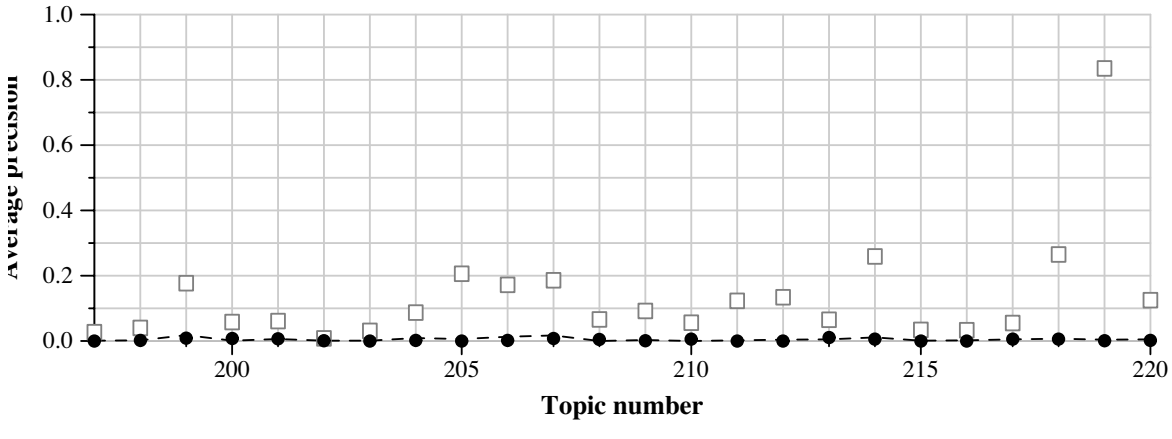


Figure 9: UIowa07AS04

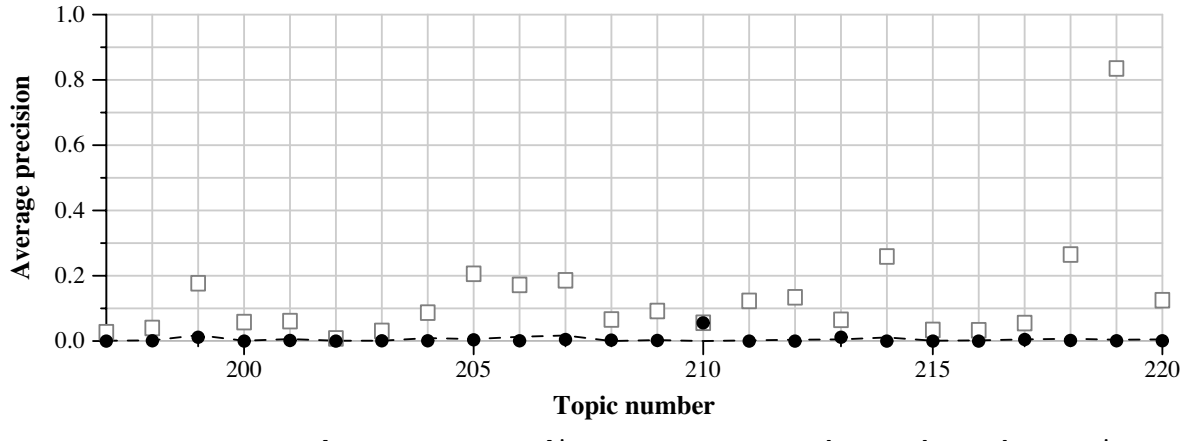


Figure 10: UIowa07AS05

References

[1] S. Brandt, J. Laaksonen, and E. Oja, Statistical shape features in content-based image retrieval, International Conference on Pattern Recognition (ICPR), 2000.

[2] S. Livens, P. Scheunders, G. Van de Wouwer, and D. Van Dyck, Wavelets for texture analysis, 6th International Conference on Image Processing and its Applications (IPA), 1997.

[3] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722.

[4] M. A. Stricker and M. Orengo, Similarity of color images, Storage and Retrieval for Image and Video Databases (SPIE), 1995, pp. 381-392.