論文 / 著書情報
Article / Book Information

Title	Tokyo Tech's TRECVID2006 Notebook
Authors	Taichi Nakamura, Yuichi Miyamura, Koichi Shinoda, Sadaoki Furui
Citation	Proc. TRECVID Workshops, Vol. , No. , pp.
Pub. date	2006, 11

TokyoTech's TRECVID2006 Notebook

Taichi Nakamura, Yuichi Miyamura, Koichi Shinoda, Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552 Japan
{nakamura,miyamura}@ks.cs.titech.ac.jp
{shinoda,furui}@cs.titech.ac.jp

In this notebook we describe our TRECVID 2006 experiments. We TokyoTech team participated in shot boundary detection and high-level feature extraction tasks.

1 Shot Boundary Detection

Our approach to shot bondary detection uses SVMs with generic features. Using the radial kernel for SVMs, we ignore the difference among the types of gradual transitions (i.e. FOI, DIS, and OTH).

We classify shot boundaries into the following three categories.

- Cuts(CUT)
- Gradual transitions with five frames or less (Short Gradual; SG)
- Gradual transitions with more than five frames (Long Gradual; LG)

We prepare a kernel function and a feature set for each of these categories.

1.1 Cut Detection

Since shot boundaries with less than five frames are classified as "cuts" in the TRECVID evaluation, the results for SG are added to the results in CUT, and are submitted as the results for "cuts". For the cut detection, we use two linear kernel SVMs (one for CUT and the other for SG) with different feature sets. The features for a CUT-SVM are activity ratio (the ratio of "dynamic" pixels to all pixels, where each dynamic pixel has larger difference than a predetermined threshold), the opticalflow, the change in the Hue-Saturation color histogram and edge. The features for SG-HMM are the activity ratio and the change in the Hue-Saturation color histogram.

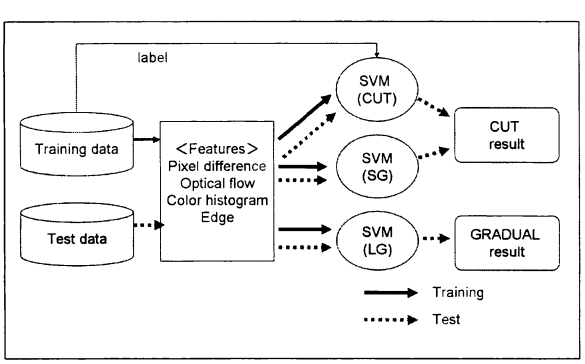

Figure 1: The shot boundary detection framework.

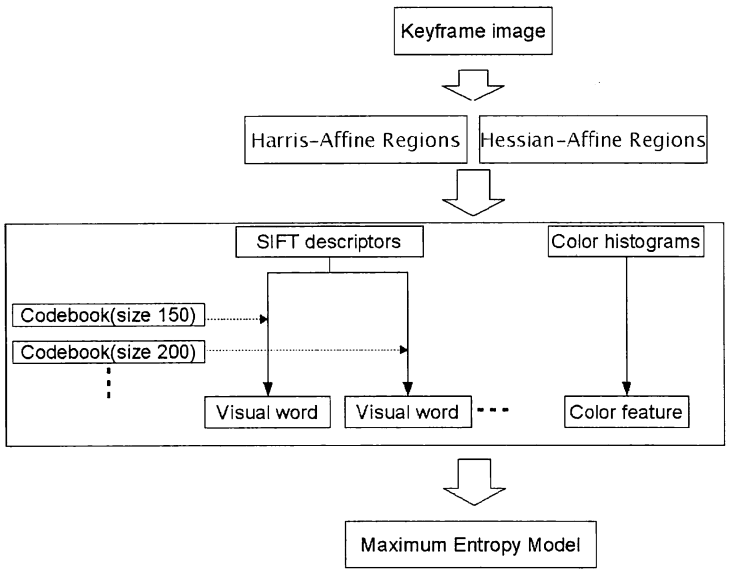

Figure 2: Overview of our visual feature extraction process

direction of the dominant gradient orientation. Finally we compute two kinds of features representing the region in the following way.

Visual words The region is first described with a 128 dimensional SIFT descriptor. This descriptor is then quantized with codebooks, which are constructed in advance for each high-level feature. We call this quantized descriptor "visual word".

Color features For each region, a 100-dimensional histogram in the Hue-Saturation color space is computed. Each region is considered to have a color feature when more than half of its pixels are in the single corresponding bin of the color histogram.

2.2 Classifiers

We use a maximum entropy model (MEM) [1] to classify the presence/absence of each of the high-level features. A MEM estimates the posterior distribution of label (presence or absence) given the features of a keyframe image. We use the implementation of MALLET [3].

2.3 Feature Selection

j Color information is effective for some high-level features, but not so effective for others. Similarly, each high-level feature has a suitable codebook size. Therefore, for each high-level feature, we tried to select the combination of feature types (e.g, color features, visual words from codebook size 150, visual words from codebook size 200, and so on) by 5 fold cross-validation. In this cross-validation, 70% of the videos were randomly selected as a training set, and remaining 30% were used for testing.

2.4 Experiments

We used the TRECVID 2005 training data set to train the MEM. The A_Tech1_2 run was trained on the complete training set using all the features. On the other hand, for A_Tech1_1 run, the feature selection technique mentioned above was used. We constructed a codebook for each high-level feature by clustering SIFT descriptors from keyframe images having the high-level feature. When there existed more than 2000 relevant keyframes, randomly sampled 2000 keyframes were used. Figure 3 shows the classification performance of our 2 runs (A_Tech1_1 and A_Tech1_2). Most of our performance is below