Visual Concept Detection

Overview

Visual concept detection allows for categorizing parts of a video for desired concept or category, e.g. *anchor shot*, *car*, *person*, etc. Within the duration of the project, a participation in the TRECVID semantic indexing competition was conducted, utilizing the technologies described here. In most current approaches to visual concept detection in videos, concept detection is performed on sampled (either regular or at certain key frames) still frames of a video. Each frame is then considered as a still image, neglecting all information about motion. In order to capture motion information, two different features have been developed.

In depth description

The baseline of our system is the standard OpponentSIFT/bag-of-words SVM classification pipeline.

The first motion feature contains of bag-of-word histograms obtained from SIFT features on the optical flow field of a frame. Therefore, optical flow is determined by using Farneback's method.

On the optical flow field, densely sampled SIFT features are extracted. This means that instead of calculating histogram of flow, histogram of gradients is chosen in order to remove the influence of global motion. A visual codebook is created with a given set of optical flow gradients by k-means clustering. Figure 21 shows the nearest neighbour assignment of the dense sampled feature points to the visual words. Hereafter, we refer to this feature as MotionSIFT.

Figure 21: Nearest neighbor assignment of feature points to visual words

The bag-of-words histogram of MotionSIFT features is concatenated with the bag-of-words histogram of key frame based OpponentSIFT features [Sande, 2008].

For the second feature multiple dimensions are observed. Therefore, we first determine feature points by dense sampling in x-, y- and t-direction. Then for each point two planes in the time domain (x-t-plane, y-t-plane) are created in addition to the x-y-plane (which is the frame itself). This is done by assembling rows or columns of subsequent frames. Figure 22 shows an example for a y-t-plane, where the columns of following frames where attached in t-direction. It is obvious that in contrast to the x-y-plane these planes contain information of motion in x- or y-direction.

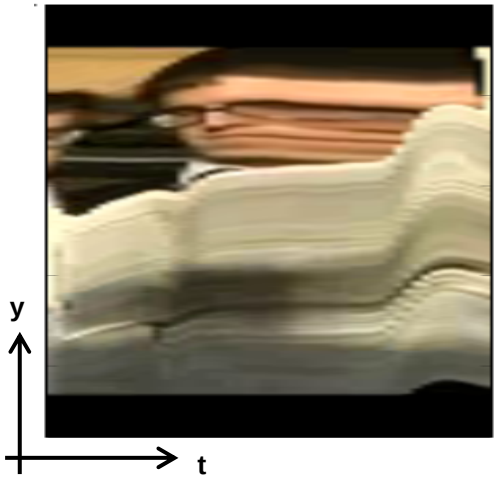

Figure 22: Example for y-t-plane (assembled columns at x=150 of subsequent frames)

Three-Dimensional SIFT features are extracted for each feature point by extracting Opponent SIFT feature for each of the three sectional planes. A visual codebook is created with a given set of three-dimensional SIFT features by k-means clustering. We use the bag-of-words histogram of three-dimensional-SIFT features for classifying without concatenated bag-of-words histograms of key frame based OpponentSIFT features, as these are already included.

Potential fields of Application

Semantic annotation of visual content can be applied in variety of usage scenarios where multimedia content annotation and retrieval is needed. It can e.g. help archivists annotating video content in order to improve re-use of existing content.

Possibilities for exploitation

Exploitation in the form of collaborative projects with potential customers is targeted.

Further Information

Further technical information is available in TOSCA-MP WP2 confidential deliverable D2.3 on “Automatic Metadata Extraction and Enrichment”.

Contact Person

Sebastian Gerke
Image Processing Department
Fraunhofer Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin, Germany