Camera Motion Detection in the Rough Indexing Paradigm

Petra Krämer, Jenny Benois-Pineau
LaBRI - CNRS, Université Bordeaux I
Domaine Univérsitaire, 351 cours de la Libération, 33405 Talence Cedex, France
{petra.kraemer,jenny.benois}@labri.fr

Abstract

This paper presents our camera motion detection method (pan, tilt and zoom) for TRECVID 2005. As input data, we only extract P-Frame motion compensation vectors directly from the MPEG compressed stream and we so achieve a performance of 3-4 times faster than real time. Our method is based on global camera motion estimation and a likelihood based significance test of the camera parameters. The best run (RI-3) on the TRECVID 2005 test set provides 0.912 for mean precision and 0.737 for mean recall.

1 Introduction

Digital videos are more and more available and pervasive due to recent progresses in storage, communication, and compression technologies. This consequently implies the increasing need for efficient indexing, browsing, search and retrieval of video archives. Requests for video material in archives often specify desired or required camera motion. Therefore, camera motion detection is a new task in TRECVID 2005. Given the feature test collection and the shot boundary reference, all shots need to be identified in which a certain camera motion (pan, tilt or zoom) is present. The videos in the test collection contain several scenes captured by a hand-carried camera. Jitter motions of the camera result. The main challenge in this task was then to overcome these jitter camera motions in order to avoid overdetections. Working on archived or broadcasted content, it is interesting to reuse motion low-level descriptors conformal in compressed streams whatever is their quality. The present work is based on our research on global (camera) motion estimation in MPEG compressed video [1].

1.1 Related Work

Hence, the fundamental problem for camera motion detection consist in estimating the camera model. Current methods for camera motion detection in MPEG compressed video generally work on motion compensation vectors or DC images extracted from the compressed stream.

The approach of Cao and Suganthan [2] is based on a neural-network scheme to characterize the camera motion in shots. They extract and reconstruct frame-by-frame motion vectors for all frames from the MPEG stream. Sàez et al. [3] estimate global motion parameters based on the Hough Transform. Their method works on DC images extracted from the compressed stream. Ewerth et al. [4] compute a 3D camera motion model only processing motion vectors from P-Frames. Due to the 3D motion model, the method allows to distinguish between translation along the x-axis (y-axis) i.e. track (boom), and rotation around the y-axis (x-axis) i.e. pan (tilt). Since in this task any distinction is supposed between track and pan or boom and tilt, these types of motion belong to the same feature groups. Ngo et al. [5] characterize camera and object motions by analysing spatio-temporal image volumes. Then, motion is depicted as oriented patterns in spatio-temporal image slices coming from DC images. They propose a tensor histogram computation in order to represent these patterns. The approach of Doulaverakis et al. [6] proposes the computation of direction histograms of MPEG motion vectors. Depending on the distribution of the histogram and the number of intracoded vectors, camera motion is detected by applying threshold values on the normalized variance of the histogram. In [7], Kim et al. present a simple thresholding scheme for the motion parameters of the affine six parameter model. The affine six parameter model is computed from MPEG motion vectors for each frame, whereas motion vectors for I-Frames are interpolated from P-Frames.

Bouthemy et al. [8] compute the 2D affine global motion model in order to characterize camera motion. Since thresholding on motion parameters is difficult mainly if jitter motions are present in the scene, thresholding is performed on likelihoods of motion parameters. This method does not work in the compressed domain. In order to evaluate their method on compressed video, they decode MPEG compressed frames. Due to its robustness, we refer to this method as a basis of our algorithm in the compressed domain.

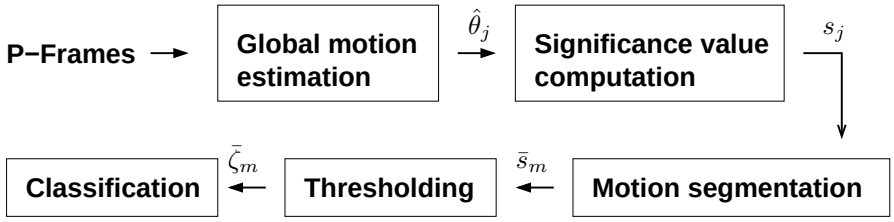


Figure 1: The steps of our algorithm for the camera motion detection in a shot. (The index j is related to frames and m to homogeneous motion segments in a shot.)

1.2 Overview of the Algorithm

We are working in the rough indexing paradigm i.e. using degraded data with a lower spatial and temporal resolution than the original stream. Thus, we use only noisy P-Frame motion compensation vectors to detect camera motion in MPEG compressed videos. Figure 1 shows an overview of our method. Our approach is based on a robust global motion estimation on noisy P-Frame motion vector fields which is the first step in this figure. We use a first order affine motion model with six parameters $\hat{\theta}$ to describe global camera motion. The global motion estimation [1] includes a motion outlier rejection scheme which handles moving objects and inaccurate encoder motion vectors for example on image borders. However, the resulting motion model parameters are still noisy due to complex motions e.g. due to a hand-carried camera. In addition, they have different meanings so that simple thresholding in order to find the dominant motion is not possible. Therefore, we chose a significance test of the motion model parameters based on [8] which is the second step in figure 1. This test is formulated as a maximisation of likelihoods s_j associated to two statistical hypotheses. One stands for the significance of the motion parameters expressing the pure physical camera motion. The second stands for the absence of the corresponding motion. Thus, the problem is turned into a better controllable problem of thresholding likelihood-ratios. We suppose that a specific motion is present in the shot if it is present in all P-Frames in a video segment of a sufficient duration. Thus, we segment a shot into video segments with homogeneous motion (step three in figure 1). To do this we consider the likelihood motion values as a stochastic signal normally distributed. Based on [8], we apply the Hinkley test on this signal allowing to detect changes on a temporal mean value $\bar{s}_m$. Due to this segmentation, the duration of a detected motion can be determined. If the duration is too short, it is considered as a jitter motion and is rejected. Then, we threshold the mean likelihood values $\bar{s}_m$ in a fourth step. Finally, a classification scheme (step five in figure 1) is applied to the thresholded mean likelihood values $\bar{\zeta}_m$ of each segment in order to define the physical character of the motion i.e. "pure" or not. We consider only segments with pure motions (pan, tilt, zoom) as a detection result. The classification using mean values eliminates subliminal jitter motions and provides the dominant motion.

Then, in the following each of these steps will be addressed. We briefly introduce the robust global motion estimation algorithm for P-Frames in section 2. The camera motion characterization is discussed in section 3. Subsection 3.1 presents the significance computation of the motion parameters, subsection 3.2 the motion segmentation method, and subsection 3.3 the classification scheme. Some results are analysed in section 4. Finally, section 5 concludes our work and outlines our points of interest for future research.

2 Global Motion Estimation from P-Frame Motion Compensation Vectors

Global motion defines the motion of the main scene content. We assume in this paper that it is principally due to the movement of the camera or the change of focus. In order to characterize global camera motion, the six-parameter affine model is used. The motion compensation vector $(dx_i, dy_i)^T$ is expressed as

$$\begin{pmatrix} dx_i \\ dy_i \end{pmatrix} = \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \quad (1)$$

where $a_1, \ldots, a_6$ are the global motion parameters of the camera. In the case of P-Frames in the MPEG compressed stream, the motion compensation vector $(dx_i, dy_i)^T$ points from the center $(x_i, y_i)^T$ of the i-th macroblock in the current image to its position in the previous image. This motion model is proved to be sufficiently rich [1][9] to characterize motion observed in an image plane of video. In addition, this TRECVID task only refers to three features groups that are pan or track, tilt or boom, and zoom or dolly. Therefore, no higher order motion model is needed to distinguish between pan and track or tilt and boom as in [4].

The method to compute global camera motion from MPEG motion fields was proposed in the previous work [1] and was tested providing good results in the shot boundary detection task of TRECVID 2004 [10]. This motion estimator follows a kind of multiresolution scheme using the weighted least square method to estimate a robust functional of motion residuals (Tukey estimator [11]).

Based on a common least square optimization problem formulation, the linear model (1) can be written in the general matrix form

$$Z = H\theta + V \tag{2}$$

where Z are the motion vectors of macroblocks, H is the observation matrix containing the macroblock centers, θ is the vector of the motion parameters and the vector V is the measurement noise.

The parameter vector can be estimated using the weighted least squares method as

$$\hat{\theta} = (H^T W H)^{-1} H^T W Z \ . \tag{3}$$

We denote N as the number of motion vectors. So, the matrices and vectors from (2) and (3) are constructed in the following way:
$\hat{\theta}$ is the 6×1 column vector of the estimated parameters from (1)

$$\hat{\theta} = (a_1, a_2, a_3, a_4, a_5, a_6)^T \ . \tag{4}$$

Z is the $2N \times 1$ column vector of the measures, in this case the MPEG compensation vectors

$$Z = (dx_1, \cdots dx_N, dy_1, \cdots dy_N)^T \ . \tag{5}$$

H is the $2N \times 6$ observation matrix

$$H = \begin{pmatrix} 1 & x_1 & y_1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & y_N & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_N & y_N \end{pmatrix} \ . \tag{6}$$

W is the $2N \times 2N$ diagonal matrix of the weights

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 & 0 \\ 0 & w_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{2N-1} & 0 \\ 0 & 0 & \cdots & 0 & w_{2N} \end{pmatrix} \tag{7}$$

where w_i are the weights calculated by the Tukey function derivation. The weight computation is based on the residuals r_i

$$r_i = z_i - \hat{z}_i \tag{8}$$

where z_i is the i-th measurement i.e. the MPEG motion compensation vector and $\hat{z}_i$ its estimation calculated with (1) using the estimated model $\hat{\theta}$.

The Tukey function ρ and its derivative ψ are defined as [1]

$$\rho(r, \lambda_r) = \begin{cases} \frac{r^6}{6} - \frac{\lambda_r^2 r^4}{2} + \frac{\lambda_r^4 r^2}{2} & \text{if } |r| < \lambda_r \\ \frac{\lambda_r^6}{6} & \text{otherwise} \end{cases} \tag{9}$$

$$\psi(r, \lambda_r) = \begin{cases} r(r^2 - \lambda_r^2)^2 & \text{if } |r| < \lambda_r \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

with λ_r as a threshold.

Then, the weight for the i-th measurement is

$$w_i = \frac{\psi(r_i)}{r_i} \ . \tag{11}$$

The weights express the adequacy of the measures to the model and thus allow for the classification of the measures as "conformant" to the model or "outliers". Then, according to the weights all macroblocks in the

current P-Frame are classified as "conformant" and "outlier" macroblocks. The latter contain object macroblocks and occluding areas. This allows to separate camera motion from object motion since macroblocks belonging to moving objects are not considered for the global motion estimation. The pure camera motion results. The subset of the conformant macroblocks will be denoted in the following as the dominant estimation support D. For details on the robust motion estimator, we refer the readers to [1].

3 Camera Motion Characterization

The global motion model parameters from the last section are still noisy due to complex motions e.g. due to a hand-carried camera. This is often the case in the TRECVID 2005 video data. In addition, the parameters in the affine model have different meanings so that simple thresholding in order to find the dominant motion is difficult. Therefore, we present a significance test of the motion model parameters based on [8].

3.1 Significance Value Computation of the Motion Parameters

Since we are interested in dominant camera motion, we express the vector of motion parameters $\theta = (a_1, ..., a_6)^T$ in another basis of elementary motion-subfields as in [8]

$$\begin{array}{l} \phi = (pan, tilt, zoom, rot, hyp1, hyp2) \quad \text{with} \\ zoom = \frac{1}{2}(a_2 + a_6) \quad rot = \frac{1}{2}(a_5 - a_3) \\ hyp1 = \frac{1}{2}(a_2 - a_6) \quad hyp2 = \frac{1}{2}(a_3 + a_5) \end{array} . \tag{12}$$

This basis is more convenient for the interpretation of the dominant motion in the scene since its is more related to the physical meaning. In [7] the affine model is transformed into a similar basis, but they consider only one hyperbolic term defined as a combination of $hyp1$ and $hyp2$.

If the dominant motion is for example a pure panning, the parameter pan is supposed to be the only non zero. This is the same for $tilt$, $zoom$ and rot. If the camera is static all parameters are supposed to be zero. In practice, this is never the case due to noise, estimation errors or moving objects. In addition, the physical meaning of the parameters is different. The values of pan and $tilt$ denote the number of pixels, while the others represent ratios. Taking into account the rich variety of camera motion in video content, it is difficult to propose an appropriate thresholding scheme to decide if a motion feature is present or not.

The significance test from [8] is a statistical approach based on a likelihood ratio test which turns the problem of direct thresholding into a better controllable problem of thresholding likelihood ratios. Two competing hypotheses are considered. The first hypothesis H_0 assumes that the considered component of ϕ is significant. The second one H_1 assumes that the component is not significant i.e. it equals zero and lets the other five parameters free. Let $\hat{\phi}_0$ and $\hat{\phi}_1$ be respectively the motion models corresponding to the hypotheses H_0 and H_1. The advantage of such a test is that it is independent from the values of the other parameters which remain free.

The likelihood function f for each hypothesis is defined with respect to the residuals $r_i = (r_{x,i}, r_{y,i})^T$ of equation (8). They are supposed to be independent, and to follow a zero-mean Gaussian law. The covariance matrices Σ_l corresponding to the two hypotheses are a posteriori estimated as

$$\Sigma_l = \begin{pmatrix} \sigma_{x,l}^2 & 0 \\ 0 & \sigma_{y,l}^2 \end{pmatrix} \tag{13}$$

with

$$\sigma_{m,l}^2 = \frac{1}{||D||} \sum_{i \in D} r_{m,i}(\hat{\phi}_l)^2, \; m = x, y \tag{14}$$

where $||D||$ is the size of the dominant estimation support D.

The two likelihood functions f for the optimized values of the motion parameters $\hat{\phi}_l$ are given by

$$\begin{aligned} f(\hat{\phi}_l) &= \prod_{i \in D} \left(\frac{1}{2\pi\sqrt{\det(\Sigma_l)}} \exp\left(-\frac{1}{2}(r_i^T \Sigma_l^{-1} r_i) \right) \right) = \frac{1}{(2\pi\sigma_{x,l}\sigma_{y,l})^{||D||}} \exp\left(-\frac{1}{2} \sum_{i \in D} (r_i^T \Sigma_l^{-1} r_i) \right) \\ &= \frac{1}{(2\pi\sigma_{x,l}\sigma_{y,l})^{||D||}} \exp\left(-||D||\right), \; l = 0, 1 \; . \end{aligned} \tag{15}$$

In order to estimate the five free parameters for hypothesis H_1, we can profit from the previous estimation of $\hat{\theta}$ i.e. $\hat{\phi}_0$ which already furnishes D. Then, a least square estimation similarly to (3) can be used. The observation

matrix H_d corresponding to the parameter vector ϕ is the $2N \times 6$ matrix

$$H_d = \begin{pmatrix} 1 & 0 & x_1 & -y_1 & x_1 & y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_N & -y_N & x_N & y_N \\ 0 & 1 & y_1 & x_1 & -y_1 & x_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & y_N & x_N & -y_N & x_N \end{pmatrix} . \tag{16}$$

In order to set the i-th component to zero in the parameter vector ϕ, we suppress in practice the i-th column in the observation matrix H_d. Thus, the least square equation is

$$\tilde{\phi} = (\mathcal{H}_d^T W \mathcal{H}_d)^{-1} \mathcal{H}_d^T W Z \tag{17}$$

where $\mathcal{H}_d$ is the reduced observation matrix H_d and $\tilde{\phi}$ is the reduced parameter vector which contains the free five parameters. The matrices W and Z remain the same than in (5) and (7).

The ratio s is called the significance value

$$\begin{aligned} s &= \ln\left(\frac{f(\hat{\phi}_1)}{f(\hat{\phi}_0)}\right) = \ln\left(\frac{\frac{1}{(2\pi\sigma_{x,1}\sigma_{y,1})^{||D||}}\exp\left(-||D||\right)}{\frac{1}{(2\pi\sigma_{x,0}\sigma_{y,0})^{||D||}}\exp\left(-||D||\right)}\right) \\ &= ||D||\left(\ln(\sigma_{x,0}\sigma_{y,0}) - \ln(\sigma_{x,1}\sigma_{y,1})\right) . \end{aligned} \tag{18}$$

We assume that $\sigma_x = \sigma_y$. Then, s becomes

$$s = ||D||\left(\ln(\sigma_0^2) - \ln(\sigma_1^2)\right) \tag{19}$$

where σ_l^2 is computed on the amplitude of the residuals r_i.

We now aim to use this value for testing the significance of the parameters. Therefore our idea is if a motion feature is present in a shot, its corresponding motion parameter is significant during a sufficient number of frames. We can not directly use the significance values for the following reasons. First of all, they can be noisy due to jitter motions. Therefore, we try to smooth them along the time and take the decision on the temporal mean of the significance values. Based on this mean, we will segment a shot into subshots of homogeneous motion. In order to get a temporal regularity, we extrapolate the significance values I-Frames. Assuming linear and constant motion, for each I-Frame the significance values of the preceding P-Frame are repeated. This seems very simplified and will be a point of future research.

The second source of noise in the significance values are the failures of the MPEG encoder and it so furnishes inaccurate motion vectors. These vectors are considered as outliers by the robust motion estimator and camera motion will be estimated only on a small part of a frame. Thus, we decide to exclude the frames where the estimation support is too small. To do this we introduce the confidence measure c_D

$$c_D = \frac{||D||}{\mathcal{D}_{max}} \tag{20}$$

with $\mathcal{D}_{max}$ as the maximum number of the estimation support i.e. the number of macroblocks in a frame. If c_D is lower than the threshold λ_D, then the motion estimated on a given P-Frame will not be further considered.

It might also be that the global motion estimation algorithm fails. In order to control the accuracy of the motion model with respect to the MPEG motion vectors, the variance of the residuals (8) computed on D are used as a second confidence measure c_σ. If c_σ exceeds the threshold λ_σ, the motion estimated on this frame is not further considered.

Then, to decide which hypothesis is selected, the mean significance value $\bar{s}$ is computed on a homogeneous motion segment M (excluding the frames rejected by the confidence measures c_D and c_σ) and the following likelihood log-test is performed

$$\bar{s} = \frac{1}{||M||}\sum_{s \in M} s \begin{array}{c} H_0 \\ < \\ > \\ H_1 \end{array} \lambda_s \tag{21}$$

with $||M||$ as the size of M. If the mean significance value $\bar{s}$ is lower than the absolute threshold λ_s, the component at hand is declared to be significant, otherwise it is considered to be null.

It is obvious that not only one component exceeds the threshold λ_s because motion in the scene is mostly a combination of basic motions. It is still possible that one dominant motion exists and though the omission of its parameter causes a much more higher increase of the error than in the case of the other remaining significant parameters. Therefore, we retain only the components who exceed λ_s and $\alpha \cdot \min\{s_{pan}, s_{tilt}, s_{zoom}, s_{rot}, s_{hyp1}, s_{hyp2}\}$[1].

3.2 Motion Segmentation

In this section we describe the method for segmenting shots into sequences of homogeneous motion. We assume that the likelihood motion values s form a stochastic signal that is normally distributed. Based on [8], we apply the Hinkley test to the signal allowing to detect changes on a temporal mean value. These changes delimit the borders of homogeneous motion segments.

Two tests are performed in parallel to look for downwards or upwards jumps. They are respectively defined by

$$U_k = \sum_{t=0}^{k} \left(s_t - \bar{s} + \frac{\delta_{min}}{2} \right) \ (k \geq 0) \tag{22}$$

$$M_k = \max_{0 \leq i \leq k} U_i; \text{ detection if } M_k - U_k > \lambda_H \tag{23}$$

$$V_k = \sum_{t=0}^{k} \left(s_t - \bar{s} - \frac{\delta_{min}}{2} \right) \ (k \geq 0) \tag{24}$$

$$N_k = \min_{0 \leq i \leq k} V_i; \text{ detection if } V_k - N_k > \lambda_H \tag{25}$$

where $\bar{s}$ is the online mean significance value before the jump defined in equation (21), δ_{min} is the minimal jump magnitude that we want to detect, and λ_H is a predefined threshold. We perform this test simultaneously on all mean significance values $(\bar{s}_{pan}, \bar{s}_{tilt}, \bar{s}_{zoom}, \bar{s}_{rot}, \bar{s}_{hyp1}, \bar{s}_{hyp2})$. If a jump has been detected on one of the signals, the means $\bar{s}$ are re-initialized for each signal.

This segmentation allows to know the duration of a certain camera motion. If the duration is too short, the segment is considered to represent jitter motion and is rejected.

3.3 Camera Motion Classification

If the segments of homogeneous motion are known, finally a classification scheme can be applied to the thresholded mean significance values $\bar{\zeta}$ of each segment in order to define the physical character of the motion. We consider only segments with pure motions (pan, tilt, zoom) as a detection result. The classification using mean values eliminates subliminal jitter motions and provides the dominant motion.

	$\bar{\zeta}$	camera motion
1	$(0,0,0,0,0,0)$	static camera/ no significant motion
2	$(\bar{\zeta}_{pan},0,0,0,0,0)$	pan
3	$(0,\bar{\zeta}_{tilt},0,0,0,0)$	tilt
4	$(\bar{\zeta}_{pan},\bar{\zeta}_{tilt},\bar{\zeta}_{zoom},0,0,0)$	zoom
5	others	complex camera motion

Table 1: Classification scheme

Table 1 shows the classification scheme we used for TRECVID to detect the physical meaning of the set of thresholded mean significance values $\bar{\zeta} = (\bar{\zeta}_{pan}, \bar{\zeta}_{tilt}, \bar{\zeta}_{zoom}, \bar{\zeta}_{rot}, \bar{\zeta}_{hyp1}, \bar{\zeta}_{hyp2})$. If a motion segment of a shot with a sufficient long duration is classified in one of the classes 2, 3 or 4, then the shot is identified to contain the corresponding motion. Since a zoom is often combined with a small pan or tilt, it is possible that the pan or tilt parameters are significant as well. This is also due to inaccurate MPEG motion vectors.

Finally, if successive segments are labelled with the same motion, the segments are joined. Two segments labelled with the same motion are joined as well if they are separated by a rejected segment.

[1] If a parameter is significant, it causes a high negative value of s.

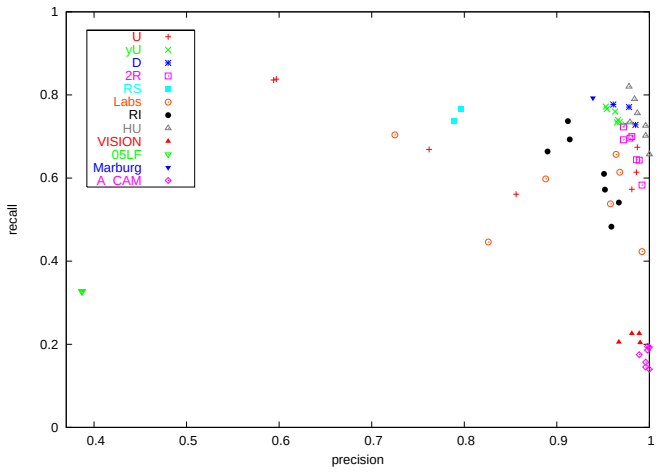


Figure 2: Precision and recall results for the submissions of all participants in the TRECVID 2005 camera motion detection task.

4 Results

For TRECVID 2005, we have several parameters to manipulate. Since no annotation is available for the development set, we annotated only a few videos for the training. In addition, if manually annotating it is difficult to decide if the feature is clearly true or not and if the annotator of the ground truth of the test set will decide in the same manner.

The parameters we use for camera motion characterization are:

- δ_{min} denotes the minimum jump magnitude that we want to detect in the Hinkley test.
- λ_H is the peak validation threshold for the Hinkley test.
- λ_s is the absolute threshold for the significance values.
- α is the constant for the relative thresholding of the significance values.
- t_{min} denotes the minimal motion duration i.e. the minimal number of frames in a valid homogeneous motion segment.
- λ_D is the absolute threshold for the confidence measure c_D i.e. the minimum size of the estimation support D.
- λ_σ is the absolute threshold for the confidence measure c_σ i.e. the accepted maximum variance of the residuals $r_i(\hat{\phi}_0)$.

The most equilibrated result (RI-3: 0.912 mean precision and 0.737 mean recall) and the best result for recall as well was obtained for the following parametrization: $\delta_{min} = 100$, $\alpha = 0.3$, $\lambda_s = -30$, $\lambda_H = 0.1$, $t_{min} = 15$, $\lambda_\sigma = 1000000$, and $\lambda_D = 0.01$. The thresholds λ_σ is chosen quite high and λ_D quite low in order to not reject too much frames. The best result for precision (RI-2: 0.967 mean precision and 0.541 mean recall) are obtained for $\delta_{min} = 100$, $\alpha = 0.25$, $\lambda_s = -70$, $\lambda_H = 0.1$, $t_{min} = 25$, $\lambda_\sigma = 1000000$, and $\lambda_D = 0.01$.

Figure 2 shows the precision and recall results for the submissions of all participants in this TRECVID task. The submission results of the group LaBRI are the black points denoted as "RI" in the key.

Figure 3 shows some results obtained in the run RI-3. It visualizes the graphs of the motion model parameters $\hat{\theta}$, the corresponding significance values s and the online mean significance values $\bar{s}$ of the shot labelled as "shot106_136". This shot is captured by a hand-carried camera and so contains a lot of jitter motions. The black lines (dotted and solid) in the graphs indicate the borders of the homogeneous motion segments. The motion segments we obtain after the joining of neighboured similar motions and the rejection of too short motion segments are marked with solid lines. Note that in this shot only the motion segments at the beginning and the end of the shot have been rejected as too short i.e. jitter motion. The camera features we detect in this shot are pan, static camera/ no significant motion, tilt and zoom. The real camera motion is a pan left followed by a zoom in. Both are correctly detected and are visualized in figure 4. A lot of jitter motion is present between these two camera motions. One part is correctly labelled as static camera or non significant motion. The other part is falsely

detected as a tilt. The graphs in figure 3(a) show the motion parameters which are very noisy. Here no zoom is visible, since the parameters indicating zoom (a_2 and a_6) have a different meaning than the parameters a_1 and a_4 respectively responsible for pan and tilt. If the significance values of the motion parameters are computed, the motions and mainly the zoom become more clear. However the graphs of the significances values in figure 3(b) are still quite noisy. This improves after the mean value computation which is shown in figure 3(c).

5 Conclusion and Perspectives

In this paper we proposed a method for camera motion detection (pan, tilt and zoom). It is based on global motion estimation and the significance test of the motion parameters without decoding the compressed stream. Only P-Frame motion compensation vectors are extracted, which allows for a fast performance i.e. 3-4 times faster than real time. The proposed method can handle moving objects in the scene and camera jitter motions.

Since no ground truth was available for the development set, it was difficult to determine the best parameter set which will work well on the test set. Then, the first point of future work will be to determine the best parameter set on the TRECVID 2005 test set. On the other hand future work will concern the improvement of our method. We will mainly focus on the correction of motion models coming from completely inaccurate P-Frame motion compensation vectors in the case if the encoder block-matching algorithm fails.

References

[1] M. Durik and J. Benois-Pineau. Robust motion characterisation for video indexing based on MPEG2 optical flow. In *International Workshop on Content-Based Multimedia Indexing, CBMI'01*, pages 57–64, 2001.

[2] X. Cao and P.N. Suganthan. Video shot motion characterization based on hierarchical overlapped growing neural gas networks. *Multimedia Systems*, 9(4):378–385, 2003.

[3] E. Saez, J.M. Palomares, J.I. Benavides, and N. Guil. Global motion estimation algorithm for video segmentation. In *Visual Communications and Image Processing, VCIP'03*, pages 1540–1550, 2003.

[4] R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben. Estimation of arbitrary camera motion in MPEG videos. In *IEEE International Conference on Pattern Recognition, ICPR'04*, volume 1, pages 512–515, 2004.

[5] C.W. Ngo, T.C. Pong, H.J. Zhang, and R.T. Chin. Motion characterization by temporal slices analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'00*, volume 2, pages 768–773, 2000.

[6] C. Doulaverakis, V. Vagionitis, M. Zervakis, and E. Petrakis. Adaptive methods for motion characterization and segmentation of MPEG compressed frame sequences. In *International Conference on Image Analysis and Recognition, ICIAR'04*, volume 1, pages 310–317, 2004.

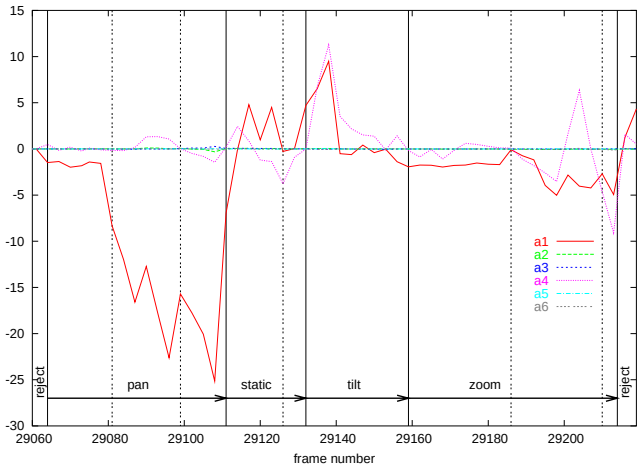
[7] J.-G. Kim, H.S. Chang, J. Kim, and H.-M. Kim. Threshold-based camera motion characterization of MPEG video. *ETRI Journal*, 26(3):269–272, 2004.

[8] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(7):1030–1044, October 1999.

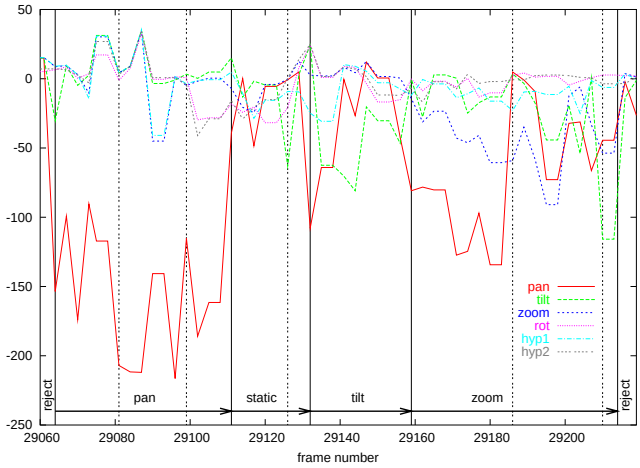
[9] Y.P. Tan, S.R. Kulkarni, and P.J. Ramadge. A new method for camera motion parameter estimation. In *IEEE International Conference on Image Processing, ICIP'95*, volume 1, pages 406–409, 1995.

[10] L. Primaux, J. Benois-Pineau, P. Krämer, and J.-P. Domenger. Shot boundary detection in the framework of rough indexing paradigm. In *TREC Video Retrieval Evaluation Online Proceedings, TRECVID'04*, 2004.

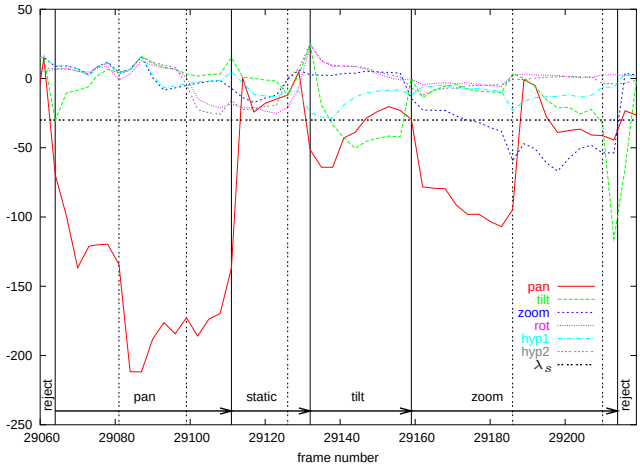
[11] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.



(a)



(b)



(c)

Figure 3: 3(a), 3(b), and 3(c) show respectively the graphs of the estimated affine global motion parameters $\hat{\theta}$, the corresponding significance values s and the online mean significance values $\bar{s}$ for the shot labelled as "shot106_136".

(a)

(b)

Figure 4: 4(a) and 4(b) show respectively the first, an intermediary and the last image for the pan and zoom correctly detected in figure 3.