14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

CAMERA MOTION CLASSIFICATION BASED ON TRANSFERABLE BELIEF MODEL

Mickael Guironnet, Denis Pellerin, and Michèle Rombaut

Laboratoire des Images et des Signaux
46 Avenue Felix Viallet, 38031, Grenoble, France
email: firstname.lastname@lis.inpg.fr

ABSTRACT

This article presents a new method of camera motion classification based on Transferable Belief Model (TBM). It consists in locating in a video the motions of translation and zoom, and the absence of camera motion (i.e static camera). The classification process is based on a rule-based system that is divided into three stages. From a parametric motion model, the first stage consists in combining data to obtain frame-level belief masses on camera motions. To ensure the temporal coherence of motions, a filtering of belief masses according to TBM is achieved. The second stage carries out a separation between static and dynamic frames. In the third stage, a temporal integration allows the motion to be studied on a set of frames and to preserve only those with significant magnitude and duration. Then, a more detailed description of each motion is given. Experimental results obtained show the effectiveness of the method.

1. INTRODUCTION

During this decade, the volume of videos has increased with the growth of diffusion processes and storage devices. To facilitate access to information, various indexing techniques using low-level features have been developed to represent video content.

Among the different features, camera motion is an important index to take into account for video content analysis and can be used in many applications such as shot segmentation [1], video summary [2] or sports video classification [3]. In general, the dominant motion is assumed to come from camera motion. A parametric model is often used to represent this, and the parameters are estimated either in compressed domain [4] or in uncompressed domain [2]. Other methods obtain camera motions by directly analyzing the MPEG motion vectors [5, 6]. However, most approaches associate a camera motion type from parameters extracted locally (either between two successive frames or from predicted pictures in MPEG) by using a learning algorithm [5] or a strategy of thresholding [4, 6]. A stage of filtering is sometimes added to obtain consistent motions [6]. But, few methods quantify identified motions. For example, a zoom is detected but the enlargement is not defined, which can be a disadvantage for some applications.

As an alternative to the various approaches presented above, we propose an original method of camera motion classification based on Transferable Belief Model (TBM). This theory is a structure adapted to process imprecise data, to combine various sources of information and to manage the conflict between the sources. The objective of our classification is to label a video in a robust way following the three camera motion classes which are: translation (pan and/or tilt), zoom and static camera. From a parametric motion model, our approach estimates frame-level camera motion, then analyzes segment-level camera motion (on a set of frames). The main contribution of this paper resides in motion recognition that is based on a certain number of rules: combination designed to avoid low magnitude motions, a filtering according to TBM to ensure the temporal coherence of the motions, and analysis on segment level to preserve the motions with consequent magnitude and duration.

The rest of the paper is organized as follows. Section 2 presents an overview of the system architecture for camera motion characterization. Section 3 discusses motion parameter extraction. After a brief description of the TBM in section 4, the method of camera motion classification is detailed in section 5. We explain in section 6 how identified motions are described. Experimental results are given in Section 7. Finally Section 8 draws our conclusions.

2. SYSTEM OVERVIEW

The system architecture is depicted in figure 1 and consists of three phases: motion parameter extraction, camera motion classification and motion description. The core of our work is the classification phase which is divided into three stages. The first stage is designed to convert the motion model parameters into symbolic values. This representation aims at facilitating the definition of rules to combine data and to provide frame-level mass functions on different camera motions. A filtering of mass functions according to TBM is carried out and contributes to ensuring the temporal coherence of the belief masses. The second stage carries out a separation between static and dynamic (zoom, translation) frames. Finally, in the third stage, the temporal integration of motions is achieved and allows motions to be studied on segment level (by gathering frames having a certain belief in a type of motion). The advantage of this analysis is to preserve only the motions with significant magnitude and duration. The description phase is then carried out by extracting different features on each video segment containing an identified camera motion type.

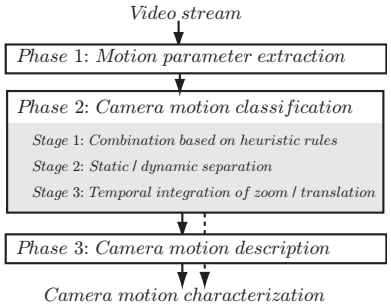

Figure 1: System architecture for camera motion characterization

3. MOTION PARAMETER EXTRACTION

The dominant motion, supposed to come from camera motion, is estimated between two successive frames by a parametric model. The affine model is chosen and can describe 5 traditional types of camera motion: zoom, rotation, horizontal translation, vertical translation, static camera. The velocity vector field is expressed according to pixel position $p_i = (x_i, y_i)$ of the frame $I(p_i, t)$ according to the following equation:

$$V_x(p_i) = c_1 + a_1 \cdot x_i + a_2 \cdot y_i$$
$$V_y(p_i) = c_2 + a_3 \cdot x_i + a_4 \cdot y_i$$

where $\theta_t = (c_1, c_2, a_1, a_2, a_3, a_4)$ are the parameters to be estimated. The determination of the model coefficients is carried out by the

Motion2D software [7]. It yields a robust and incremental estimation of the dominant motion exploiting the spatio-temporal derivatives of the frame intensity.

Before to use these coefficients, we achieve an average filter of size L_1 on the parameters θ_t to reduce noise and estimation errors. Certain model parameters are specific to a motion and are used to identify camera motions. From the filtered parameters $\theta'_t = (c'_1, c'_2, a'_1, a'_2, a'_3, a'_4)$, the displacement of the camera $dpl(t)$ and the divergent $div(t)$ between two successive frames $I(p_i,t)$ and $I(p_i,t+1)$ are defined as well as the total displacement $dpt(t_o,t_f)$ and the distance traveled $dtl(t_o,t_f)$ between two times t_o and t_f:

$$\begin{array}{lll} \overrightarrow{dpl}(t) & = (c'_1(t), c'_2(t)) & dpt(t_o,t_f) = \left\| \sum_{j=t_o}^{t_f-1} \overrightarrow{dpl}(j) \right\| \\ dpl(t) & = \left\| \overrightarrow{dpl}(t) \right\| & \\ div(t) & = (a'_1(t) + a'_4(t))/2 & dtl(t_o,t_f) = \sum_{j=t_o}^{t_f-1} \left\| \overrightarrow{dpl}(j) \right\| \end{array}$$

The total displacement dpt corresponds to the displacement in the straight line between the original and final position whereas the distance traveled dtl is the original way and corresponds to the integration of all displacements between sampling times.

According to the magnitude of the variables div and dpl, the different camera motions can be extracted. A translation (respectively a zoom) is detected if the displacement (respectively the divergent) is high. When a light zoom and a strong translation occur simultaneously, the zoom is not, or hardly visible and thus should not be detected. In the same way, only the zoom is preserved in the presence of a strong zoom and a weak translation. In order to satisfy these rules, the variables need to be converted into linguistic values to be combined. Before describing camera motion classification, the following section will point out the bases of the Transferable Belief Model.

4. TRANSFERABLE BELIEF MODEL

The Transferable Belief Model was formalized by P. Smets [8] and comes from the Dempster-Shafer's evidence theory.

Let $\Omega = \{H_1, \cdots, H_N\}$ be the frame of discernment containing N mutually exclusive and exhaustive hypotheses related to a given problem. A mass function or a Basic Belief Assignment (BBA) is a function $m : 2^\Omega \rightarrow [0,1]$ that assigns a value in [0,1] to each subset A of Ω. The value $m(A)$ is the part of belief that is allocated exactly to the proposition A. Under closed-world assumption, a BBA is subject to the following constraints: $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$. The subsets $A \subseteq \Omega$ such that $m(A) > 0$ are called focal elements of m.

Consider two BBA m_1 and m_2 defined on the same frame of discernment and provided by a source 1 and a source 2 respectively. According to applications, two combinations are possible: conjunctive combination $m_1 \textcircled{\cap} m_2(A_i)$ and disjunctive combination $m_1 \textcircled{\cup} m_2(A_i)$.

$$\begin{array}{l} m_1 \textcircled{\cap} m_2(A_i) = \sum_{A_j \cap A_k = A_i} m_1(A_j) \cdot m_2(A_k) \\ m_1 \textcircled{\cup} m_2(A_i) = \sum_{A_j \cup A_k = A_i} m_1(A_j) \cdot m_2(A_k) \end{array}$$

The conjunctive combination (respectively disjunctive) is interpreted as a logical "and" (respectively "or"). These combinations can then be used in logical rules. From a BBA, a transformation was proposed by P. Smets [8] to obtain a probability measure called pignistic probability on the frame of discernment Ω:

$$BetP^\Omega(A) = \sum_{B \subseteq \Omega} \frac{m^\Omega(B)}{1 - m^\Omega(\emptyset)} \frac{|A \cap B|}{|A|}, \ \forall A \subseteq \Omega \qquad (1)$$

where $|A|$ is the cardinal of $A \subseteq \Omega$. This function can be used for decision-making.

Let Ω_1 and Ω_2 be two distinct and disjointed frames of discernment, a BBA can be defined on $\Omega = \Omega_1 \times \Omega_2$ through the conjunctive combination as follows:

$$m_1 \textcircled{\cap} m_2(A \times B) = m_1(A) \cdot m_2(B) \ \forall A \subseteq \Omega_1, \ \forall B \subseteq \Omega_2$$

The interest of the Cartesian product is to apply TBM even when the frames of discernment are disjointed and thus not compatible.

5. CAMERA MOTION CLASSIFICATION

Camera motion classification consists in locating in a video the places where a camera motion takes place. The method, which is based on TBM, has to identify the three camera motions that are translation, zoom and absence of motion. The principle of camera motion classification phase is presented in figure 2. It is divided into three stages: combination based on heuristic rules, static/dynamic separation and temporal integration of zoom and translation motions. The rest of this section describes each stage of the method.

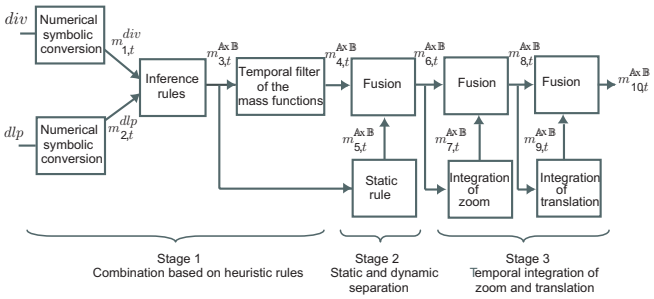

Figure 2: Principle of camera motion classification phase

5.1 Combination based on heuristic rules

The first stage (fig. 2) consists in converting the model parameters into symbolic values describing the retrieved motions. From these variables, we establish heuristic rules to combine them in order to affect frame-level belief masses on camera motions. Then a temporal filtering of belief masses is carried out for the purpose of ensuring the temporal coherence of the belief masses on a neighborhood.

5.1.1 Numerical-symbolic conversion

The numerical variables dpl and div are transformed into symbolic values: weak (W), average (A), large (L) and very large (VL). A type of fuzzy sets is used to formalize expert knowledge and to provide a symbolic representation of data. Each linguistic term is associated to a set defined by a function as indicated in figure 3. With regard to the symbolic description of divergent, it is carried from the absolute value of divergent that informs about the magnitude of the zoom (the sign is related to the direction). Thus the mass functions for the variables div and dpl are respectively defined on the frame of discernment $div = \{W,A,L,VL\}$ and $dpl = \{W,A,L,VL\}$. The combination of these mass functions will lead to camera motion detection.

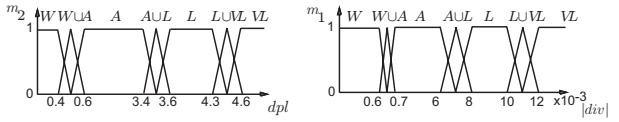

Figure 3: Definition of the BBAs for the displacement (left) and for the divergent (right)

5.1.2 Inference rules

The approach to camera motion classification is based on heuristic rules. The Transferable Belief Model (TBM) provides tools adapted to build models that integrate inference mechanisms.

Let $\mathbb{A} = \{T, \overline{T}\}$ be the frame of discernment of the translation motion and let $\mathbb{B} = \{Z, \overline{Z}\}$ be the frame of discernment of the zoom motion where T (resp Z) is a hypothesis on the presence of the translation (resp zoom) and $\overline{T}$ (resp $\overline{Z}$) is an hypothesis on the absence

of the translation (resp absence of zoom). The motion identification can be carried out by applying the Cartesian product of sets $\mathbb{A} \times \mathbb{B}$. For example, if a frame belongs to class $(\overline{T},\overline{Z})$ then the frame is regarded as static. The rules R which we defined for camera motion classification are summarized in table 1. For example, if div is average and dpl is large then the detected motion is translation and thus the belief mass on the product set $\mathbb{A} \times \mathbb{B}$ is assigned to $(T,\overline{Z})$. We can also notice propositions on several motions such as $(\overline{T},\overline{Z}),(\overline{T},Z)$. This means the absence of translation and ignorance of the presence of zoom, which corresponds to a proposition on "static or zoom". The combination m_1^{div} and m_2^{dpl} with the rules R leads to the definition of a new BBA $m_3^{\mathbb{A}\times\mathbb{B}} = m_1^{div} \textcircled{R} m_2^{dpl}$ which directly characterizes the belief on camera motions. We can also note that the rules are designed to avoid as far as possible secondary motions. For example, if the displacement is very large and the divergent is large then the zoom motion is neglected and only the translation motion is considered. Finally, a BBA $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ is obtained for each frame t of the video.

		div			
		Weak	*Average*	*Large*	*Very Large*
dpl	*Weak*	$(\overline{T},\overline{Z})$	$(\overline{T},\overline{Z}),(\overline{T},Z)$	$(\overline{T},Z)$	$(\overline{T},Z)$
	Average	$(\overline{T},\overline{Z}),(T,\overline{Z})$	$\mathbb{A}\times\mathbb{B}$	$(\overline{T},Z)$	$(\overline{T},Z)$
	Large	$(T,\overline{Z})$	$(T,\overline{Z})$	(T,Z)	$(\overline{T},Z)$
	Very Large	$(T,\overline{Z})$	$(T,\overline{Z})$	$(T,\overline{Z})$	(T,Z)

Table 1: Attribution of a BBA in function of the values of divergent and displacement according to rules R

5.1.3 Temporal filtering of mass functions

Temporal filtering of mass functions was introduced and is based on the hypothesis that camera motion cannot be very different from one frame to the next. If the case appears, then it is considered that all motions are possible, without being able to highlight one rather than another. This filtering according to TBM, adds doubt by reallocating the belief on the union of motion propositions if the temporally close beliefs deliver different information. The filter is produced by the disjunctive combination of the sources $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ on a temporal window of size L_2. A new BBA is then obtained:

$$m_{4,t}^{\mathbb{A}\times\mathbb{B}} = m_{3,t-(L_2-1)/2}^{\mathbb{A}\times\mathbb{B}} \textcircled{\cup} \cdots \textcircled{\cup} m_{3,t+(L_2-1)/2}^{\mathbb{A}\times\mathbb{B}}$$

The interest of this combination is to increase the temporal coherence of motions and thus prevent the presence of different motions on a neighborhood. The consistency of a motion can be improved by filling the possible holes generated by estimation errors.

Figure 5 shows an example of sequence having a zoom out where the method is applied with a window of size $L_1 = L_2 = 13$. When the divergent is average and the displacement is weak, the motion is considered to be "static or zoom" $(\overline{T},\overline{Z}),(\overline{T},Z)$ on curve $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$. When the displacement becomes average with an average divergent, the mass is allocated to total doubt $\mathbb{A} \times \mathbb{B}$. The temporal filtering (curve $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$) amplifies the zone of $\mathbb{A} \times \mathbb{B}$ by adding doubt. The two following stages of classification allow camera motion to be found .

Globally, the rules and the filtering correspond to a very cautious process leaving a wide place open to doubt rather than imposing a final opinion on camera motion.

5.2 Static/dynamic separation

The second stage (fig. 2) consists in separating the static frames from the dynamic frames (zoom, translation) by taking into account the temporal neighborhood of beliefs allocated locally by the heuristic rules (here the preceding filtering is not considered). In the absence of camera motion, the estimated model parameters have often a weak magnitude. However this property is not always checked locally because of noise or estimation errors. To take it into account, a frame will be considered as static if the close frames are static. Thus a new BBA is defined and is based on the following rule: if a certain number of frames around the frame studied have a belief on the static hypothesis (respectively dynamic) then a belief mass will be allocated to the static hypothesis (respectively dynamic) for the frame studied.

Let $\Omega = \{S,D\}$ be the frame of discernment where S and D indicate a static and dynamic motion respectively. Ω is a coarsening of $\mathbb{A} \times \mathbb{B}$ and reciprocally $\mathbb{A} \times \mathbb{B}$ is a refinement of Ω. To know if the frame t is rather static or rather dynamic, each BBA $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ is transformed into a BBA $m_{3,t}^{\Omega}$ as follows:

$$\begin{aligned} m_{3,t}^{\Omega}(S) &= m_{3,t}^{\mathbb{A}\times\mathbb{B}}(\overline{T},\overline{Z}) \\ m_{3,t}^{\Omega}(D) &= \textstyle\sum_{K\subseteq\mathbb{A}\times\mathbb{B}\setminus(\overline{T},\overline{Z})} m_{3,t}^{\mathbb{A}\times\mathbb{B}}(K) \\ m_{3,t}^{\Omega}(\Omega) &= 1 - m_{3,t}^{\Omega}(S) - m_{3,t}^{\Omega}(D) \end{aligned}$$

For each frame t, a temporal window of size L_3 centered on t is considered. The proportion of frames having a static hypothesis on this window is determined by combining the BBA m_t^{Ω} on the Cartesian product $\Omega' = \Omega_{t-(L_3-1)/2} \times \ldots \times \Omega_{t+(L_3-1)/2}$ where each Ω_i is associated to frame i of window centered on frame t studied. If a mass of Cartesian product Ω' has at least $\alpha\%$ of frames on the static hypothesis then this mass is deferred to the static hypothesis S for the frame t studied. In the same way, if a mass of Ω' has at least $100-\alpha\%$ of frames on the dynamic hypothesis then it is affected to the dynamic hypothesis D for the frame t studied. If it is not the case, then the mass is returned to the hypothesis $S \cup D$.

Based on this rule, a BBA $m_{5,t}^{\Omega}$ on Ω is defined for each frame t. It is extended to $\mathbb{A} \times \mathbb{B}$ to be combined with $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$ using the conjunctive combination. The resulting BBA is $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ and if the mass attributed to the empty set is non-null then it is transferred to the union of propositions.

In figure 5, as the motion is either "static or zoom" $(\overline{T},\overline{Z}),(\overline{T},Z)$ or total doubt $\mathbb{A} \times \mathbb{B}$, the separation cannot find static camera since no mass is associated to the proposition "static".

5.3 Temporal integration of zoom and translation

The third stage (fig. 2) achieves a more global motion description on segment level (by gathering frames containing the same motion type). This consists in segmenting the sequence by coherent motions (translation or zoom), then estimating the motion magnitude on each segment. By describing motion on each segment, the purpose of this integration is to preserve only motions of consequent magnitude and duration.

5.3.1 Case of zoom

As soon as the pignistic probability $BetP^{\mathbb{A}\times\mathbb{B}}(\{(\overline{T},Z),(T,Z)\})$ on a frame (eq. 1) becomes higher than a threshold δ then the beginning of a zoom is detected and the corresponded time t_0 is memorized. When $BetP^{\mathbb{A}\times\mathbb{B}}(\{(\overline{T},Z),(T,Z)\})$ is lower than δ then the zoom motion stops and this time t_f is memorized. The segment between t_0 and t_f contains a potential zoom which is analyzed to be ensured of its presence. As the divergent is not very well adapted to represent zoom, the enlargement coefficient is introduced.

We develop the case in one dimension. Let $a'_1(t)$ be the parameter of the affine model at the time t and let v_x be the velocity for the position x_i provided by $v_x = a'_1(t) \cdot x_i$ assuming the other coefficients to be null (case for a perfect zoom). The position at the time $t+1$ is given by $x'_i = x_i + v_x = x_i \cdot (1 + a'_1(t))$. From where the report between the position at the final time t_f and the position at the initial time t_0 is given by:

$$k_x = \prod_{t=t_0}^{t_f-1} (1 + a'_1(t))$$

If the motion is a zoom in, the ratio k_x corresponds to an enlargement coefficient ($k_x > 1$), denoted ag_x. If it is a zoom out then k_x

is a reduction coefficient ($k_x < 1$) and by convention the inverse of this report $ag_x = 1/k_x$ is called enlargement coefficient. In the case of a frame, an enlargement coefficient ag is defined by multiplying ag_x and ag_y obtained along the axis x and y. The enlargement coefficient ag represents the ratio between frame size and the part of the frame that increased until frame size. For zoom segment, the sign of divergent can change, which means a change of zoom direction. In order to take this into account, we determine on each zoom segment, the under-segments having the divergent of the same sign and an enlargement is calculated on each one of them.

Finally, the enlargement coefficient ag that characterizes the power of the zoom is used to cancel or preserve the zoom segment. Thus, a BBA $m_7^{\Omega_Z}$ is built on the frame of discernment $\Omega_Z = \{\overline{Zoom}, Zoom\}$ from the enlargement coefficient as shown in figure 4. $m_7^{\Omega_Z}$ is then extended on $\mathbb{A}\times\mathbb{B}$ using the relations $\{(\overline{T},Z),(T,Z)\} = Zoom$, $\{(\overline{T},\overline{Z}),(T,\overline{Z})\} = \overline{Zoom}$ and $\mathbb{A}\times\mathbb{B} = \Omega_Z$ and the resulting BBA $m_7^{\mathbb{A}\times\mathbb{B}}$ is associated to each frame of the segment. The passage of description from segment to frame allows the BBA $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ defined previously to be combined with this one and the resulting BBA is $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$. Table 2 shows the combination of the masses. It is important to note that, in case of conflict, $m_{7,t}^{\mathbb{A}\times\mathbb{B}}$, being more reliable than $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ for the "zoom", the mass associated to the empty set is transfered to the proposition of the zoom coming from $m_{7,t}^{\mathbb{A}\times\mathbb{B}}$ and to the proposition of the translation coming from $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$.

		$m_{6,t}^{\mathbb{A}\times\mathbb{B}}$			
		$(\overline{T},\overline{Z})$	$(T,\overline{Z})$	$(\overline{T},Z)$	(T,Z)
$m_7^{\mathbb{A}\times\mathbb{B}}$	$(\overline{T},\overline{Z}),(T,\overline{Z})$	$(\overline{T},\overline{Z})$	$(T,\overline{Z})$	$\emptyset \to (\overline{T},\overline{Z})$	$\emptyset \to (T,\overline{Z})$
	$(\overline{T},Z),(T,Z)$	$\emptyset \to (\overline{T},Z)$	$\emptyset \to (T,Z)$	$(\overline{T},Z)$	(T,Z)
	$\mathbb{A}\times\mathbb{B}$	$(\overline{T},\overline{Z})$	$(T,\overline{Z})$	$(\overline{T},Z)$	(T,Z)

Table 2: Combination of the mass functions $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ and $m_7^{\mathbb{A}\times\mathbb{B}}$

5.3.2 *Case of translation*

As processing with zoom, a segment of potential translation is obtained using $BetP^{\mathbb{A}\times\mathbb{B}}(\{(T,\overline{Z}),(T,Z)\}) > \delta$, then the segment between t_o and t_f is analyzed by calculating maximum displacement dpl_{max} on this window.

$$t = \arg\max_{t_k\in[t_o,t_f]}(dpl(t_o,t_k)) \quad and \quad dpl_{max} = dpl(t_o,t)$$

Maximum displacement dpl_{max} is then standardized by the duration (from t_o to t) to have a relative representation of displacement. Thus standardized maximum displacement dpl_{maxn} characterizes the power of translation on the segment and this value is used to define a mass function (fig. 4) on $\Omega_T = \{Translation, \overline{Translation}\}$. Like the zoom, $m_9^{\Omega_T}$ is extended on $\mathbb{A}\times\mathbb{B}$, then this one is associated to each frame of segment to be combined with $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$ and the resulting BBA is $m_{10,t}^{\mathbb{A}\times\mathbb{B}}$.

Figure 4: Definition of the BBAs from the enlargement coefficient (left) and the standardized maximum displacement (right)

The integration is applied with $\delta = 0.1$ in figure 5. We can see that the integration of the zoom (curve $m_{7,t}^{\Omega_Z}$) allows it to be preserved and thus leads to the removal of the "static" proposition on the curve $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$. Then, the integration of the translation (curve $m_{9,t}^{\Omega_T}$) allows it to be removed and thus only the proposition $(\overline{T},Z)$ on the curve $m_{10,t}^{\mathbb{A}\times\mathbb{B}}$ is preserved.

Finally the decision on camera motions is taken by choosing the maximum of the pignistic probability for each frame.

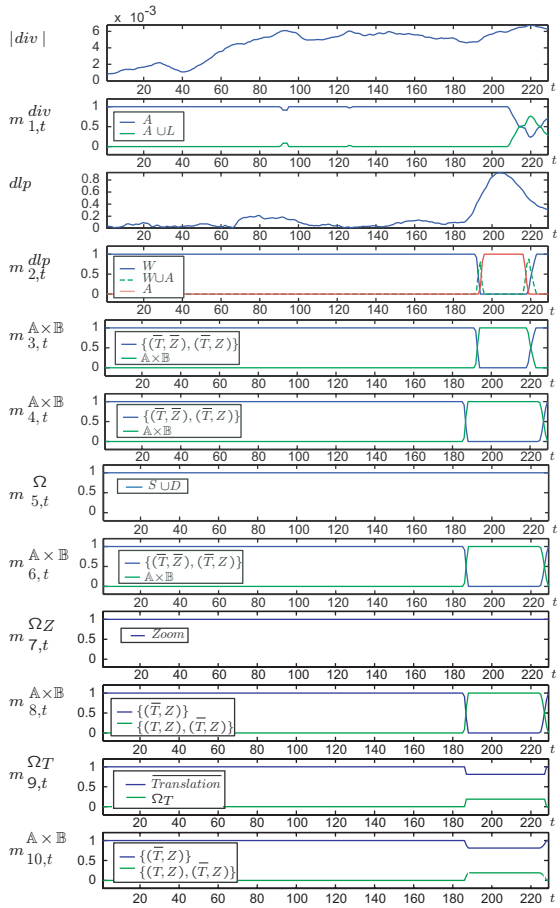

Figure 5: Illustration of the classification method on a sequence having a motion of zoom out. The awaited motion is $(\overline{T},Z)$.

6. CAMERA MOTION DESCRIPTION

This phase (fig. 1) consists in describing each identified camera motion. For the three motions (static, translation and zoom), a binary decision is attributed to each frame. Based on the results of the previous paragraphs, each segment where a zoom is identified is described by the enlargement coefficient and the direction. The sign of the divergent is used to know the zoom direction (zoom in or zoom out). The translation segment is represented by distance traveled and standardized total displacement. Moreover, the translation direction is also obtained for each frame contained in a translation segment. A fuzzy quantification (fig. 6) from vector phase $\overrightarrow{dpl}(t)$ is used to represent it. For example, a diagonal motion from down-left to up-right is characterized by the four values $(Zone\ 1, Zone\ 2, Zone\ 3, Zone\ 4) = (0.5, 0.5, 0, 0)$.

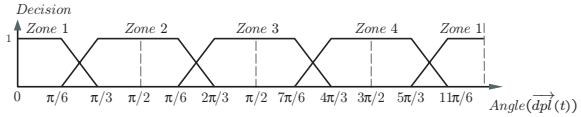

Figure 6: Membership functions according to the 4 directions.

7. CAMERA MOTION CLASSIFICATION EVALUATION

Camera motion classification evaluation aims to verify the performance of the method. Two studies are discussed: one on video extracts containing a single camera motion and an other containing composed camera motions. Thereafter, we apply the method with the following thresholds: $L_1 = L_2 = L_3 = 13$ (window about a half second), $\alpha = 50\%$ (stage 2) and $\delta = 0.1$ (stage 3).

7.1 Analysis of single motions

To evaluate the method of camera motion classification, video extracts containing a single camera motion type were selected during the video playback. The chosen video extracts are various contents (sport sequences, series "The Avengers",...) and possess perceived motions. The corpus is composed of 42 video extracts (4605 frames) of a few seconds each (8 extracts at fixed camera, 21 extracts containing a translation, and 13 extracts containing a zoom).

The results are reported for motion classification (presence of static, translation or zoom). Like evaluation measures, we use recall and precision. However, the classification of a video extract depends on camera motion allocated on each one of these frames. We consider that a video is correctly identified if all frames are correctly classified. Table 3 shows the results of motion classification. If the zoom is considered, recall indicates that one video is not found. It is about a zoom which is in fact detected at 73%. The beginning of this video has a light zoom and is related with a static camera. With regard to the translation motion, it misses one video for the recall, this one is detected at 95% and has a small static segment at the beginning. Hence camera motion classification presents good performances with a precision of 100%, a recall $> 92\%$ for the three camera motions, which demonstrates the robustness of the method.

	Translation	Zoom	Static
Recall	*95.2 (20/21)*	*92.3 (12/13)*	*100 (8/8)*
Precision	*100 (20/20)*	*100 (12/12)*	*100 (8/8)*

Table 3: Performance of the classification of video extracts

Table 4 illustrates the description of zoom and translation according to the direction. Here, a video is correctly identified if at least 80% of frames are well classified. The obtained results shows the performance of the motion direction description.

	Right to left	Up to down	Left to right	Down to up	Zoom in	Zoom out
Recall at 80%	*100 (6/6)*	*100 (7/7)*	*100 (6/6)*	*100 (2/2)*	*100 (7/7)*	*83.3 (5/6)*
Precision at 80%	*100 (6/6)*	*100 (7/7)*	*100 (6/6)*	*100 (2/2)*	*100 (7/7)*	*100 (5/5)*

Table 4: Local performance of the classification of video extracts with recall and precision calculated at 80%

7.2 Analysis of composed motions

Camera motion classification is studied here on video extracts where the motions can be superimposed (zoom and translation) or successive in the same extract. We annotated three video extracts (a documentary of sports with 20 shots and 3271 frames, a series "The Avengers" with 27 shots and 2412 frames, and TV news with 42 shots and 6870 frames) according to the three camera motions. By assuming the known shots, the different motions are extracted by the method and compared with the ground truth. The evaluation is carried out by recall and precision on frame level (calculation of the frame number correctly identified for each motion). Nevertheless, the ground truth is sometimes difficult to determine in certain places of the video (ambiguity between motions) or the border between two successive motions is difficult to find. From these considerations, errors can be added to the classification errors coming from the classification method. That allows the results presented in table 5 to be moderated. We can notice that the results are good with a recall and a precision superior to 79% for the three videos.

	documentary	News	series
Recall	*84.8 (2884/3401)*	*85.9 (5990/6967)*	*90.0 (2409/2663)*
Precision	*79.4 (2884/3632)*	*85.0 (5990/7045)*	*88.6 (2409/2718)*

Table 5: Performance of the classification of frames on three video extracts with composed motions

8. CONCLUSION

We have presented a method of camera motion classification based on Transferable Belief Model. It consists in finding the motions of translation and zoom, and static camera in a video. The approach is characterized by its rule based recognition system. The combination rules are designed to avoid as far as possible secondary motions (low magnitude motions). A filtering according to TBM is carried out and modifies the belief in a motion following the close frames. A static/dynamic separation is archived and assumes that a frame is static if these close frames are considered to be static. Finally the analysis on segment level aims at only preserving motions of consequent magnitude and duration. Then, the description of motions is carried out by quantifying them (for example, coefficient enlargement for a zoom) to interpret them easily. The motion of translation and zoom are also characterized in a more local way with the direction (zoom in, zoom out, translation from left to right...).

In order to ensure the performances of the method, we have presented results on videos containing one motion type or containing superimposed camera motions or which followed one another. In the two cases, the results obtained in term of recall and precision enable us to conclude that our classifier is effective to determine camera motions. One of the future line of investigation would be to consider other motion types such as rotation. TBM will not pose a theoretical problem to integrate them.

REFERENCES

[1] Y. Qi, A. Hauptmann, and T. Liu, "Sports video categorizing method using camera motion parameters," in *Proc. of ICME*, vol. 2, Baltimore, USA, 6-9 July 2003, pp. 689–692.

[2] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler, "A geometrical key-frame selection method exploiting dominant motion estimation in video," in *Proc. of CIVR*, Dublin, Ireland, 21-23 July 2004, pp. 419–427.

[3] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga, "Sports video categorizing method using camera motion parameters," in *VCIP*, Lugano, Switzerland, 2003.

[4] J.-G. Kim, H. S. Chang, J. Kim, , and H.-M. Kim, "Threshold-based camera motion characterization of mpeg video," *ETRI Journal*, vol. 26, no. 3, pp. 269–272, June 2004.

[5] C. Chen, C. Bhumireddy, and P. K. Darvemula, "Camera motion classification using a genetic functional-link neural network," in *Proc. of International Conference on IROS*, vol. 3, Sandai, Japon, 28 Sept - 2 Oct 2004, pp. 2343 – 2348.

[6] X. Zhu, A. Elmagarmid, and A. C. Catlin, "Insightvideo: Toward hierarchical content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.

[7] J. M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of VCIR*, vol. 6, no. 4, pp. 348–365, Dec. 1995.

[8] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, Dec. 1994.