HAL
open science

A fusion architecture based on TBM for camera motion classification

Mickaël Guironnet, Denis Pellerin, Michèle Rombaut

To cite this version:

Mickaël Guironnet, Denis Pellerin, Michèle Rombaut. A fusion architecture based on TBM for camera motion classification. Image and Vision Computing, 2007, 25 (11), pp.1737-1747. 10.1016/j.imavis.2007.01.001 . hal-00164601

HAL Id: hal-00164601

https://hal.science/hal-00164601v1

Submitted on 21 Jul 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A fusion architecture based on TBM for camera motion classification

M. Guironnet [a], D. Pellerin [a,*], M. Rombaut [a]

[a] *Grenoble Image Parole Signal Automatique (GIPSA-lab) (ex. LIS)*
46 Avenue Felix Viallet, 38031 Grenoble, France

Abstract

We propose in this paper an original 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.

Key words: camera motion classification, Transferable Belief Model, motion estimation, motion description, video indexing

* Corresponding author. Tel.: +33476574369; Fax:+33476574790.
Email addresses: `mickael.guironnet@yahoo.fr` (M. Guironnet), `denis.pellerin@lis.inpg.fr` (D. Pellerin), `michele.rombaut@lis.inpg.fr` (M. Rombaut).
URL: `http://www.gipsa-lab.inpg.fr`.

1 INTRODUCTION

Recently, the volume of videos has increased spectacularly with the growth of storage devices and progress of diffusion processes. Consequently this proliferation of video sequences has led to the emergence of new applications such as video summary, classification or browsing in a video base. It has hence become necessary to index video content efficiently to facilitate access to information. Generally, video description relies on the extraction of low-level features such as color, texture or motion to which it is difficult to give a meaning. This description is thus a difficult task which has been studied in many research works.

Among the different features, camera motion is an important index to take into account for video content analysis. From camera motion, much semantic information can be deduced such as the activity in a scene. For example, an action movie contains many scenes with strong camera motions to give rhythm. Furthermore, the way of filming a scene can also direct the gaze. A zoom in will focus the attention on a precise zone of the scene. The knowledge of camera motion can also be exploited to separate moving objects from the background and can be used in the algorithms of segmentation. In this way, this index is an interesting tool to extract the semantic context of the scene.

Using camera motion classification, sport scenes could be labelled such as various cricket shots [1], types of American football moves [2] and basketball video sequences [3]. In the same way, by analyzing the statistics of camera motions, Takagi et al. [4] show that camera motion is a sufficient signature to differentiate sports activities and classify them. Camera motion can also be used in many applications such as shot segmentation [5,6], video summary [7,8] as well as in models of visual attention [9]. Moreover, requests for video material from archives have become so great that in 2005 a new task was inserted into the TREC Video [10] experiments concerning camera motion classification, considering the main objective of TREC Video is to promote progress in content-based retrieval from digital video.

Generally, 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 [11,3,12] or in uncompressed domain [7,6,8]. Other methods obtain camera motions by directly analyzing the MPEG motion vectors [13–16]. 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 [13,16] , a strategy of thresholding [11,14] or a template-matching algorithm [15]. A stage of filtering is sometimes added to obtain consistent motions [14]. However, 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 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 main camera motion classes which are: translation (pan and/or tilt), zoom and static camera. The translation corresponds either to a rotational motion of the camera about the vertical and/or horizontal axis, or to the tracking of the camera along the vertical and/or horizontal axis. The zoom leads to the enlargement or the reduction of part of the image. Lastly, static camera is a scene obtained with fixed camera where the objects can move. These three camera motions are similar to those of TREC Video 2005 with the determination of the horizontal translation (pan), the vertical translation (tilt) and the zoom. From a parametric motion model, the proposed approach estimates frame-level camera motion, then analyzes segment-level camera motion (on a set of frames). Although the motion estimation used in this work is carried out in uncompressed domain, our method can be adapted to the compressed domain as in [12,11,3]. Indeed, the model parameters which are handled can be indifferently estimated in the compressed or uncompressed domain. The main contribution of this paper resides in the motion recognition that is based on a certain number of rules: combination designed to avoid identifying low magnitude camera 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 classification and description. 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 the 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 the proposed system 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 facilitates the definition of rules to combine data and to provide frame-level mass func-

tions 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 the 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. For example, a zoom motion is characterized by a enlargement coefficient.

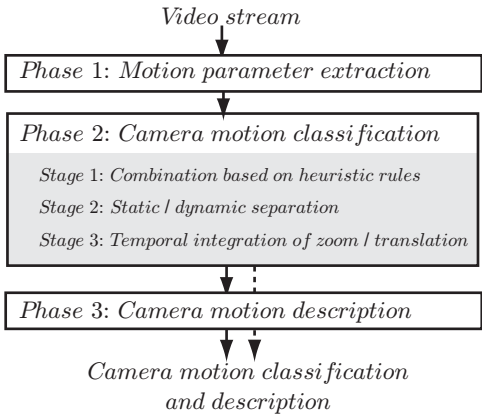

Fig. 1. System architecture for camera motion classification and description

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 for the 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 [17]. 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, an average filter of size L_1 on the parameters θ_t is achieved in order to reduce noise and estimation errors. An example of the parameter estimation is shown in figure 2 where the sequence contains a zoom in. It can be seen that the parameters are disturbed and sometimes erroneous (strong impulses between image 400 and 450). The temporal window size is chosen by expertise to 13 frames corresponding to half of a second which is coherent to the human reflex time.

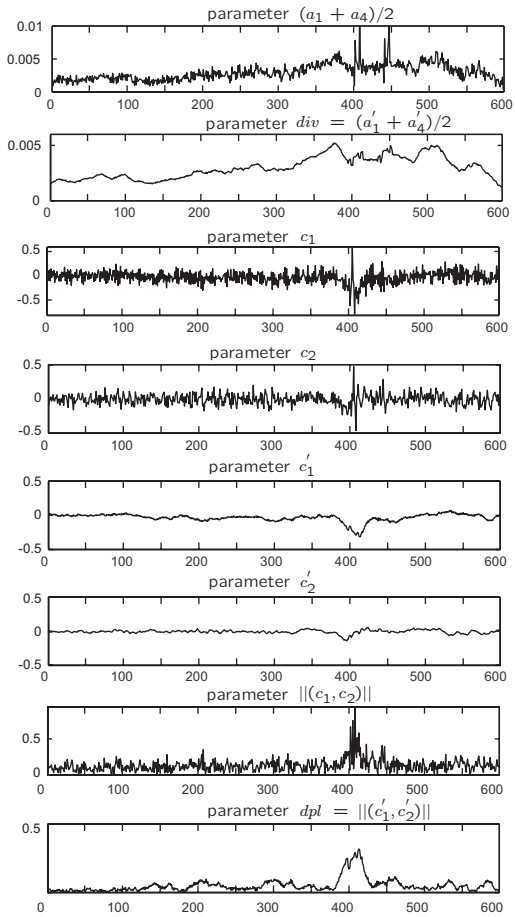

Fig. 2. The parameters (in pixels/frame) evolve according to time on a sequence containing a zoom in. As regards the zoom motion, the only parameter expected to be higher than 0 is $(a_1 + a_4)/2$ and the parameters c_1 and c_2 are expected to be null.

Some 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:

$$\begin{aligned}
\overrightarrow{dpl}(t) &= (c_1^{'}(t), c_2^{'}(t)) \\
dpl(t) &= \left\|\overrightarrow{dpl}(t)\right\| \\
div(t) &= \tfrac{1}{2}(a_1^{'}(t) + a_4^{'}(t)) \\
dpt(t_o, t_f) &= \left\|\sum_{j=t_o}^{t_f-1} \overrightarrow{dpl}(j)\right\| \\
dtl(t_o, t_f) &= \sum_{j=t_o}^{t_f-1} \left\|\overrightarrow{dpl}(j)\right\|
\end{aligned}$$

The total displacement *dpt* in pixels/frame 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 highlight. 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 [18] 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. From the frame of discernment, the power set of Ω denoted as 2^Ω is defined and is composed of all subsets of Ω (singleton and composed hypotheses).

$$2^\Omega = \{A/A \subseteq \Omega\} = \{\emptyset, \{H_1\}, \cdots, \{H_N\}, \{H_1, H_2\}, \cdots, \Omega\}$$

It is assumed that the solution to a given problem is necessarily in the frame

of discernment (closed world). On the contrary, the open-world assumption admits the existence of hypotheses not defined in the frame of discernment.

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 $m_1 \textcircled{\cup} m_2(A_i)$.

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

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 [18] to obtain a probability measure called pignistic probability on the frame of discernment Ω:

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

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 depends on TBM, has to identify the three camera motions that are translation, zoom and the absence of motion. It also has to recognize strong and short motions as well as weak and long motions, and to avoid false detections due to a poor estimation. The principle of camera motion classification phase is presented in figure 3. It is divided into three stages: combination based on heuristic rules, static/dynamic separation and temporal integration of zoom and translation motions. The sequel of this section describes each stage of the method.

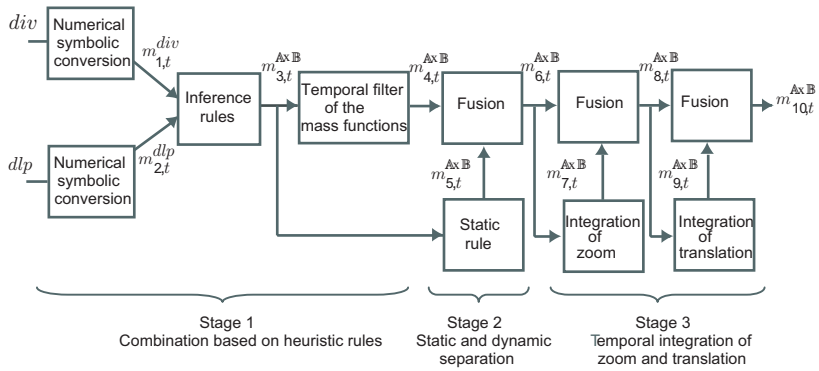

Fig. 3. Principle of camera motion classification phase

5.1 Combination based on heuristic rules (stage 1)

The first stage (fig. 3) 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 give frame-level belief masses on the different camera motions. Then a temporal filtering of belief masses is carried out for 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 or group is associated to a set defined by a function as indicated in figure 4. With regard to the symbolic description of divergent, it is carried from the absolute value of divergent (4.a). Indeed, the absolute

value only gives information about the amplitude of the zoom whereas the direction of the zoom is obtained by the sign. 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.

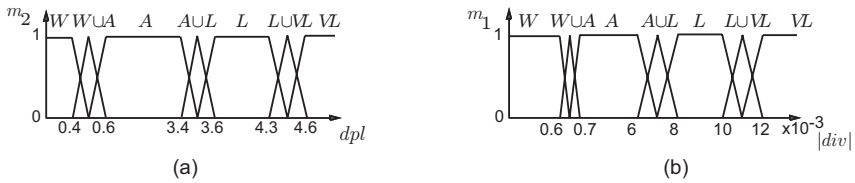

Fig. 4. Definition of the BBAs for the displacement *dpl* (a) and for the divergent *div* (b) in pixels/frame

Because it is difficult to get annotated data base, the thresholds are fixed by expertise. The TBM framework allows the expert to directly model the doubt by union of hypotheses: a large doubt avoid risk but does not give lot of information.

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. On the other hand, if the frame belongs to class (T, Z) then the frame has a camera motion at the same time of translation and zoom.

We define the set of rules R to attribute belief masses to the product set $\mathbb{A} \times \mathbb{B}$:

- If *div* is weak and *dpl* is weak then the camera motion is static $(\overline{T}, \overline{Z})$.
- If *div* is average and *dpl* is large then the motion detected is translation $(T, \overline{Z})$.
- If *div* is very large and *dpl* is very large then the motion detected is translation and zoom (T, Z).
- and so on.

These rules R defined for camera motion classification are summarized in

table 1. For example, if *div* is large and *dpl* is average then detected motion is zoom and thus the belief mass on the product set $\mathbb{A} \times \mathbb{B}$ is assigned to $(\overline{T}, Z)$. 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". In the same way, $\{(\overline{T}, \overline{Z}), (T, \overline{Z})\}$ means the absence of zoom and ignorance of the presence of translation, which corresponds to a proposition on "static or translation" whereas $\mathbb{A} \times \mathbb{B}$ is a total ignorance of the camera motion. The combination m_1^{div} and m_2^{dpl} with the rules R leads to the definition of a new BBA $m_3^{\mathbb{A}\times\mathbb{B}} = m_1^{div} \circledR m_2^{dpl}$ which directly characterizes the belief on camera motions. It can also be noticed 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 rules R according to divergent *div* and displacement *dpl*

In order to explain the method, we propose to process the example in which, for a given image, the non null masses are m_1^{div} and m_2^{dpl} as follows:

$$m_1^{div}(\{W\}) = 0.7 \qquad m_2^{dpl}(\{A, L\}) = 0.2$$
$$m_1^{div}(\{W, A\}) = 0.3 \qquad m_2^{dpl}(\{L\}) = 0.8$$

The BBA $m_3^{\mathbb{A}\times\mathbb{B}}$ has 3 focal elements:

$$\begin{aligned}
m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T}, \overline{Z}),\ (T, \overline{Z})\right\}\right) &= m_1^{div}(\{W\}) \cdot m_2^{dpl}(\{A, L\}) = 0.14 \\
m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T, \overline{Z})\right\}\right) &= m_1^{div}(\{W\}) \cdot m_2^{dpl}(\{L\}) + m_1^{div}(\{W, A\}) \cdot m_2^{dpl}(\{L\}) \\
&= 0.80 \\
m_3^{\mathbb{A}\times\mathbb{B}}(\mathbb{A} \times \mathbb{B}) &= m_1^{div}(\{W, A\}) \cdot m_2^{dpl}(\{A, L\}) = 0.06
\end{aligned}$$

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:

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

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 8-a 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 decision on camera motion.

5.2 Static/dynamic separation (stage 2)

The second stage (fig. 3) 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 majority of 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. In order 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:

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

From the previous example,

$$\begin{aligned} m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T},\overline{Z}),\ (T,\overline{Z})\right\}\right) &= 0.14 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T,\overline{Z})\right\}\right) &= 0.80 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\mathbb{A}\times\mathbb{B}\right) &= 0.06 \end{aligned}$$

mass m_3^{Ω} is obtained:

$$\begin{aligned} m_3^{\Omega}(\{S,D\}) &= m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T},\overline{Z}),\ (T,\overline{Z})\right\}\right) + m_3^{\mathbb{A}\times\mathbb{B}}\left(\mathbb{A}\times\mathbb{B}\right) = 0.20 \\ m_3^{\Omega}(\{D\}) &= m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T,\overline{Z})\right\}\right) = 0.80 \end{aligned}$$

For each frame t, a temporal window of size L_3 centered on t is considered. The new BBA for the frame t defined on Ω is deduced from the conjunctive combination of the BBAs of all the frames of the window and defined 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. The coarsening process from Ω' to Ω depends for each subset of Ω' on the number of frames that are considered to be static (S), dynamic (D) or doubt between these hypotheses. If the subset has at least $\alpha\%$ of frames on the static hypothesis, then this combination mass is deferred to the static hypothesis S for the frame t studied. In the same way, if the subset 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 set of hypotheses $\{S,D\}$.

The number of images attributing a mass to the static hypothesis is determined by $n = ceil(L_3.\alpha)$ where *ceil* means round up to nearest integer. The number of images with a mass on the dynamic hypothesis is obtained by $L_3 - n + 1$.

The value of $\alpha = 50\%$ is a compromise between the ability of detection of real static frames and the risk of false detection.

Let us take the example with $L_3 = 3$ and $\alpha = 50\%$ with $\Omega_1 = \{S_1, D_1\}$, $\Omega_2 = \{S_2, D_2\}$ et $\Omega_3 = \{S_3, D_3\}$. The number of images that must affect a

mass to the static hypothesis is $n = 2$, whereas for the dynamic hypothesis the number of images is of $L_3 - n + 1 = 2$. The belief mass distribution of image 2 is modified according to beliefs of close images (here image 1 and 3). An example is shown in equation 2 how the proposition of cartesian product Ω' is redistributed on Ω:

$$\begin{aligned}
&\{S_1\} \times \{D_2\} \times \{S_3\} && \rightarrow \{S\} \\
&\{S_1\} \times \{D_2\} \times \{D_3\} && \rightarrow \{D\} \\
&\{S_1\} \times \{D_2\} \times \{S_3, D_3\} && \rightarrow \{S, D\}
\end{aligned} \tag{2}$$

If on three successive images, there is a strong belief mass related to static hypothesis for image 1 and 3, and a strong belief mass allocated to dynamic motion for image 2:

$$\begin{aligned}
m_3^{\Omega_1}(\{S_1, D_1\}) &= 0.3 & m_3^{\Omega_1}(\{S_1\}) &= 0.7 \\
m_3^{\Omega_2}(\{S_2, D_2\}) &= 0.4 & m_3^{\Omega_2}(\{D_2\}) &= 0.6 \\
m_3^{\Omega_3}(\{S_3, D_3\}) &= 0.1 & m_3^{\Omega_3}(\{S_3\}) &= 0.9
\end{aligned}$$

The BBA m_5^{Ω} then is obtained:

$$\begin{aligned}
m_5^{\Omega}(\{S\}) &= m_3^{\Omega_1}(\{S_1\}) \cdot m_3^{\Omega_2}(\{S_2, D_2\}) \cdot m_3^{\Omega_3}(\{S_3\}) + \\
&\quad m_3^{\Omega_1}(\{S_1\}) \cdot m_3^{\Omega_2}(\{D_2\}) \cdot m_3^{\Omega_3}(\{S_3\}) \\
&= 0.63 \\
m_5^{\Omega}(\{S, D\}) &= 0.37 \\
m_5^{\Omega}(\{D\}) &= 0
\end{aligned}$$

This example shows that the distribution of the masses depends on the proportion of beliefs on the static and dynamic hypotheses. As two images out of three have a strong belief on the static hypothesis, the resulting mass supports the belief on the static hypothesis and the mass allocated to D is null.

Based on this rule, a BBA $m_{5,t}^{\Omega}$ on Ω is defined for each frame t. It is extended to $\mathbb{A} \times \mathbb{B}$ using the relations $\{(\overline{T}, \overline{Z})\} = S$, $\{(\overline{T}, Z), (T, \overline{Z}), (T, Z)\} = D$ and $\mathbb{A} \times \mathbb{B} = \Omega$, and it is 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. Table 2 recapitulates the combination of the two BBA.

		$m_{5,t}^{\mathbb{A}\times\mathbb{B}}$		
		$(\overline{T},\overline{Z})$	$(T,\overline{Z}),(\overline{T},Z),(T,Z)$	$\mathbb{A}\times\mathbb{B}$
$m_{4,t}^{\mathbb{A}\times\mathbb{B}}$	$(\overline{T},\overline{Z})$	$(\overline{T},\overline{Z})$	$\emptyset\rightarrow\mathbb{A}\times\mathbb{B}$	$(\overline{T},\overline{Z})$
	$(T,\overline{Z})$	$\emptyset\rightarrow(T,\overline{Z}),(\overline{T},\overline{Z})$	$(T,\overline{Z})$	$(T,\overline{Z})$
	$(\overline{T},Z)$	$\emptyset\rightarrow(\overline{T},Z),(\overline{T},\overline{Z})$	$(\overline{T},Z)$	$(\overline{T},Z)$
	(T,Z)	$\emptyset\rightarrow(T,Z),(\overline{T},\overline{Z})$	(T,Z)	(T,Z)

Table 2
Combination of BBA $m_{5,t}^{\mathbb{A}\times\mathbb{B}}$ and $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$ using the rule of conjunctive combination and managing the empty set.

Figure 5 illustrates the second stage of this approach with the following parameters $\alpha = 50\%$ and $L_3 = L_2 = L_1 = 13$. The sequence is filmed with fixed camera. For example, a belief mass is assigned to the proposition "zoom or statics" (between the images 85 and 105 , and images 38 and 42 of $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$) and the dynamic/static separation redistributes the belief mass to the "static" proposition on the curve $m_{5,t}^{\mathbb{A}\times\mathbb{B}}$. On the other hand, the dynamic/static separation is not sufficient to allocate the belief to the "static" proposition on all along the segment. Indeed, for example, no mass is associated to the "static" proposition between images 1 and 10 of $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ and the resulting mass (curves $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$) is assigned with the "static or dynamic" proposition. It is the integration of this segment which will find the static camera. In figure 8-b, 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".

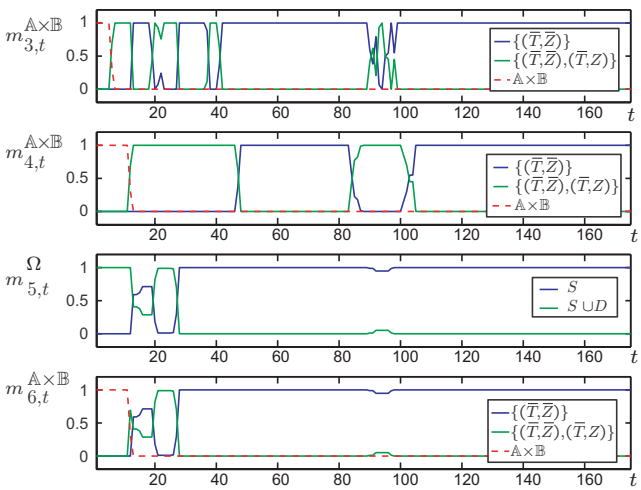

Fig. 5. Stage 2: Illustration of filtering and static/dynamic separation on a filmed sequence with fixed camera. The expected motion is static $(\overline{T},\overline{Z})$.

5.3 Temporal integration of zoom and translation (stage 3)

The third stage (fig. 3) 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 zoom is detected and this 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. In order to detect this potential motion, δ is chosen sufficiently low. 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 ratio 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. On the other hand, if it is a zoom out then k_x is a reduction coefficient ($k_x < 1$) and by convention the inverse of this ratio $ag_x = 1/k_x$ is called enlargement coefficient. In the case of a frame (2 dimensions), one enlargement coefficient ag_x is defined along the x-axis and one ag_y following the y-axis. To obtain only one enlargement coefficient ag, the two coefficients ag_x and ag_y are multiplied. 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 sub-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 on the segment or the sub-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 6.

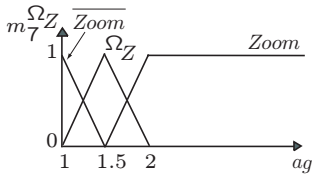

Fig. 6. Definition of BBA for the enlargement coefficient.

$m_7^{\Omega_Z}$ 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 3 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)
$m_7^{\mathbb{A}\times\mathbb{B}}$	$(\overline{T}, \overline{Z}), (T, \overline{Z})$	$(\overline{T}, \overline{Z})$	$(T, \overline{Z})$	$\emptyset \rightarrow (\overline{T}, \overline{Z})$	$\emptyset \rightarrow (T, \overline{Z})$
	$(\overline{T}, Z), (T, Z)$	$\emptyset \rightarrow (\overline{T}, Z)$	$\emptyset \rightarrow (T, Z)$	$(\overline{T}, Z)$	(T, Z)
	$\mathbb{A} \times \mathbb{B}$	$(\overline{T}, \overline{Z})$	$(T, \overline{Z})$	$(\overline{T}, Z)$	(T, Z)

Table 3
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 dpt_{max} on this window.

$$t = \underset{t_k \in [t_o, t_f]}{\arg\max}(dpt(t_o, t_k)) \text{ and } dpt_{max} = dpt(t_o, t)$$

Maximum displacement dpt_{max} is then standardized by the duration (from time t_o to t) to have a relative representation of displacement. Thus standardized maximum displacement dpt_{maxn} characterizes the power of translation on the segment and this value is used to define a mass function (fig. 7) 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 $m_{10,t}^{\mathbb{A}\times\mathbb{B}}$.

The integration is applied with $\delta = 0.1$ in figure 8-c. In fact, if δ is low, the motion detection is improved. That can lead to false alarms, but in this

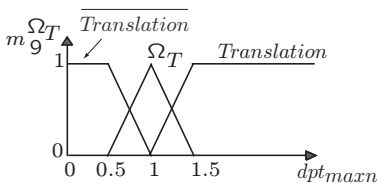

Fig. 7. Definition of the BBA for the standardized maximum displacement

algorithm, the temporal integration insures the reliability. 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 to 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.

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 obtained for each frame contained in a translation segment. A fuzzy quantification (fig. 9) 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 $(Zone\ 1, Zone\ 2, Zone\ 3, Zone\ 4) = (0.5, 0.5, 0, 0)$.

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).

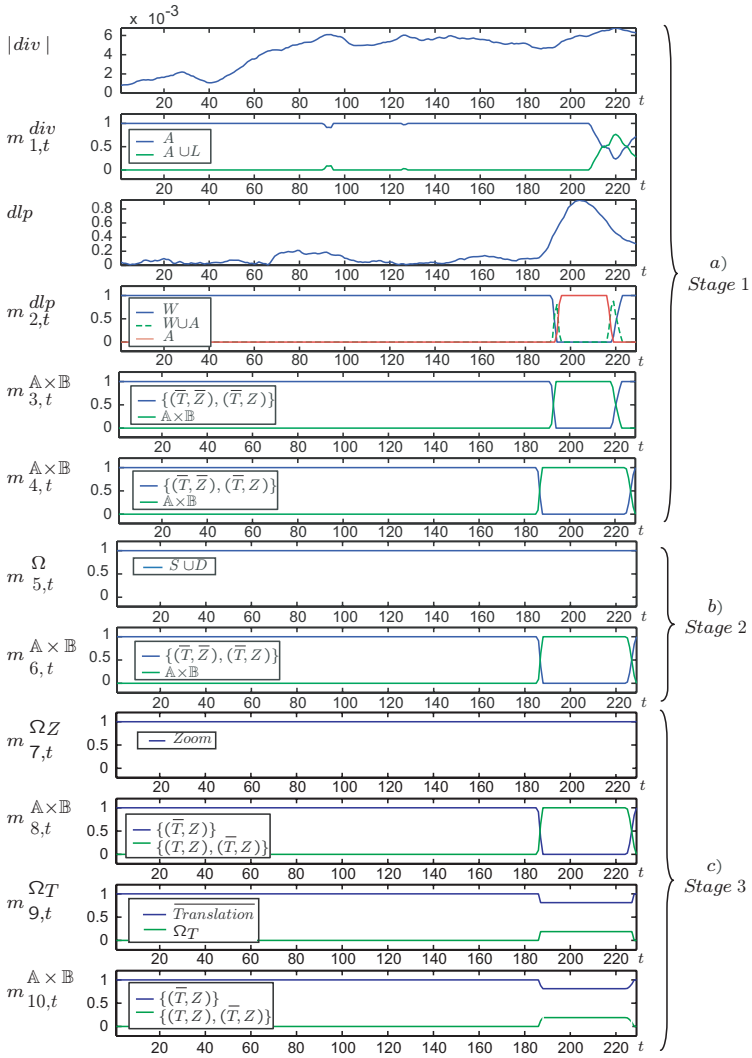

Fig. 8. Illustration of the classification method (stages 1, 2 and 3) on a sequence having a motion of zoom out. The expected motion is $(\overline{T}, Z)$. The variables *div* and *dlp* are in pixels/frame.

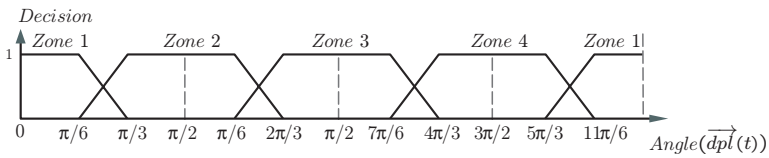

Fig. 9. Membership functions according to the 4 directions.

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 (fig.10) are various contents (sport sequences, series "The Avengers",. . .) and possess perceived motions. The corpus is made up

of 42 video extracts (4605 frames) of a few seconds each:

- 8 extracts (899 images on total) at fixed camera
- 21 extracts (2053 images on total) containing a translation (6 translations from right to left, 7 from top to bottom, 6 from left to right and 2 from bottom to top)
- 13 extracts (1663 images on total) containing a zoom (7 zooms in and 6 zooms out)

Fig. 10. Examples of video extracts contained in the base. For each example, the left image corresponds to the first image of the extract and the right image is the last image of the extract. Two examples are filmed with fixed camera (in top), two include a translation motion (in the medium) and finally two contain a zoom (in bottom).

The results are reported for motion classification (presence of static, translation or zoom). Like evaluation measures, we use recall and precision. Recall R evaluates the capacity of the classifier to find the videos in the base containing a retrieved motion and is defined as the number of relevant video extracts retrieved containing the desired motion in a database divided by the total number of video extracts retrieved. Precision P evaluates the capacity of the classifier to find only videos having the desired motion and is defined as the number of relevant video extracts retrieved containing the desired motion divided by the total number of relevant video extracts in a database. 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 4 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 (20/21)*	*92 (12/13)*	*100 (8/8)*
Precision	*100 (20/20)*	*100 (12/12)*	*100 (8/8)*

Table 4
Performance of the classification of video extracts

Table 5 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 (5/6)*
Precision at 80%	*100 (6/6)*	*100 (7/7)*	*100 (6/6)*	*100 (2/2)*	*100 (7/7)*	*100 (5/5)*

Table 5
Performance of the description of zoom and translation according to direction 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. Figure 11 shows an example of video extract including several camera motions, initially a segment of zoom out and translation of left to right, then a segment of zoom and finally a static segment. This example also underlines the camera motion description. The indicators of the different camera motions can also be found (4 directions of the translation, the distance traveled, standardized total displacement as well as the direction of the zoom and the enlargement coefficient).

To evaluate the method, we annotated three video extracts according to the three camera motions:

- a sports documentary with 20 shots and 3271 frames
- "The Avengers" series with 27 shots and 2412 frames
- TV news with 42 shots and 6870 frames

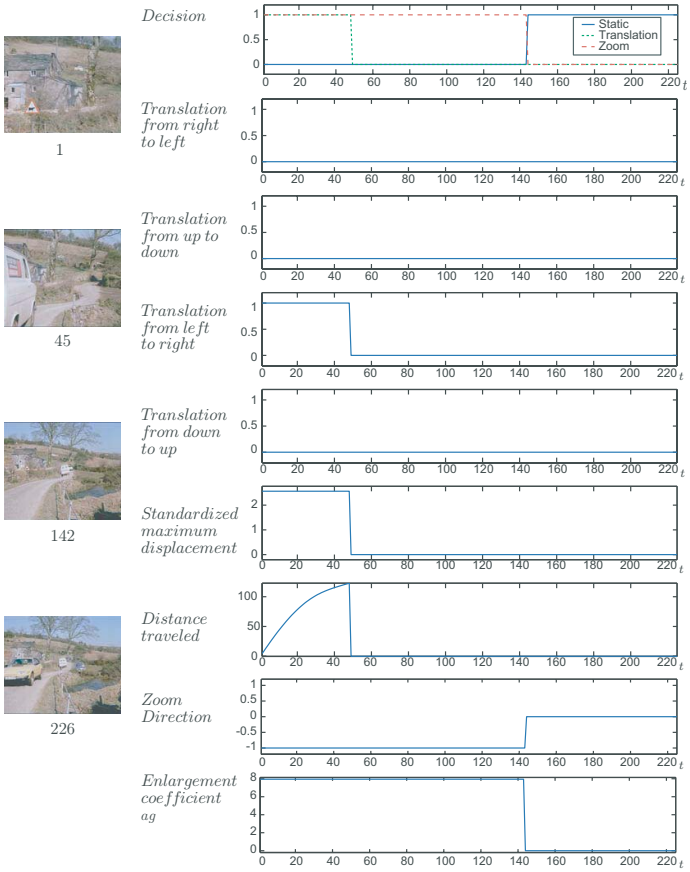

Fig. 11. Classification on an video extract including several camera motions.

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 6 to be moderated. We can note that the results are good with more than to 70% recall and precision for the three videos. With regard to static, it is dominating in the three videos and is detected with good accuracy. The motion of translation is also easily found in the videos. On the other hand, the motion of zoom is the least present in the videos. Considering this small quantity, the results are relatively good with more than to 70% recall and precision. Figure 12 is an example of the camera motion classification and corresponds to the first shot of the "The Avengers" sequence where a translation motion is followed by a static segment. We can notice that the motions identified by the method are similar to those of the ground truth. As the border between the motions is not exactly at the same place, the recall and the precision are

81% and 100% for statics and 100% and 95% for the translation whereas the motion determination seems to be correct.

		documentary	**News**	**series**
Static	**Recall**	*78 (1083/1386)*	*96 (3898/4052)*	*91 (1181/1304)*
	Precision	*97 (1083/1112)*	*84 (3898/4661)*	*96 (1181/1236)*
Translation	**Recall**	*92 (1251/1366)*	*71 (1891/2658)*	*96 (853/889)*
	Precision	*72 (1251/1734)*	*90 (1891/2098)*	*85 (853/1003)*
Zoom	**Recall**	*85 (550/649)*	*78 (201/257)*	*80 (375/470)*
	Precision	*70 (550/786)*	*70 (201/286)*	*78 (375/479)*

Table 6
Performance of the classification of frames on three video extracts with composed motions

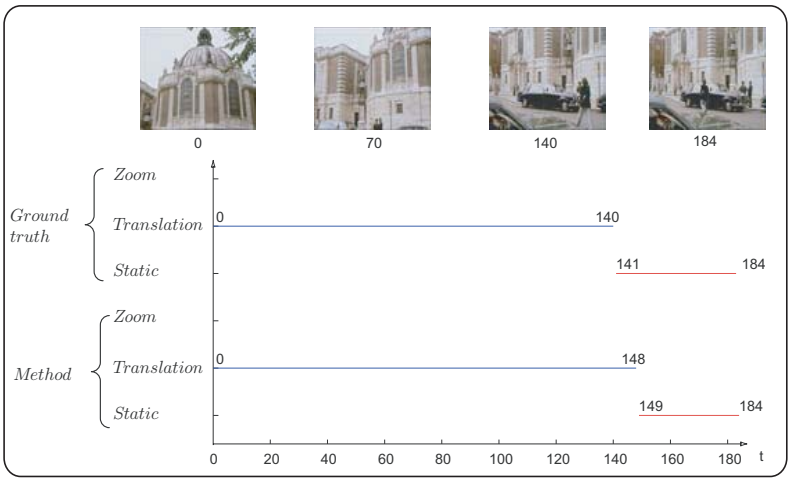

Fig. 12. Example of camera motion classification on the first shot of the "The Avengers" sequence.

8 Conclusion

A method of camera motion classification based on Transferable Belief Model has been presented. 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, enlargement coefficient 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 lines of investigation would be to consider other motion types such as rotation. The advantage of the TBM framework is that it is easy to add new parameters without changing the structure of the system. If it is assumed the rotation is relevant enough, the associated parameters could be included. Lastly, our method requires knowledge of shot change. A detector could also be designed from camera motion.

References

[1] M. Lazarescu, S. Venkatesh, G. West, On the automatic indexing of cricket using camera motion parameters, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02), Vol. 1, Laussane, Switzerland, 2002, pp. 809–813.

[2] M. Lazarescu, S. Venkatesh, Using camera motion to identify different types of american football plays, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03), Vol. 2, Baltimore, USA, 2003, pp. 181–184.

[3] Y.-P. Tan, D. Saur, S. Kulkarni, P. Ramadge, Rapid estimation of camera motion from compressed video with application to video annotation, IEEE Trans. Circuits Syst. Video Technol. 10 (1) (2000) 133–146.

[4] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, H. Tominaga, Sports video categorizing method using camera motion parameters, in: Visual Communications and Image Processing (VCIP'03), Vol. 5150, Lugano, Switzerland, 2003, pp. 2082–2088.

[5] Y. Qi, A. Hauptmann, T. Liu, Sports video categorizing method using camera motion parameters, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03), Vol. 2, Baltimore, USA, 2003, pp. 689–692.

[6] P. Bouthemy, M. Gelgon, F. Ganansia, A unified approach to shot change detection, IEEE Trans. Circuits Syst. Video Technol. 9 (1999) 1033–1044.

[7] S. V. Porter, M. Mirmehdi, B. T. Thomas, A shortest path representation for video summarisation, in: Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03), Mantova, Italy, 2003, pp. 460–465.

[8] B. Fauvet, P. Bouthemy, P. Gros, F. Spindler, A geometrical key-frame selection method exploiting dominant motion estimation in video, in: Conference on Image and Video Retrieval (CIVR'04), Dublin, Ireland, 2004, pp. 419–427.

[9] W.-H. Chend, W.-T. Chu, J.-L. Wu, A visual attention based region-of-interest determination framework for video sequences, IEICE Transactions on Information and Systems Journal E88-D (7) (2005) 1578–1586.

[10] The nist trec video retrieval evaluation (2005).
URL `http://www-nlpir.nist.gov/projects/tv2005/tv2005.html`

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

[12] W. J. Gillespie, D. T. Nguyen, Robust estimation of camera motion in mpeg domain, in: Proceedings of Conference on Analog and Digital Techniques in Electrical Engineering (TENCON'04), Vol. 1, Chiang Mai, Thailand, 2004, pp. 395–398.

[13] C. Chen, C. Bhumireddy, P. K. Darvemula, Camera motion classification using a genetic functional-link neural network, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS'04), Vol. 3, Sandai, Japon, 2004, pp. 2343 – 2348.

[14] X. Zhu, A. Elmagarmid, A. C. Catlin, Insightvideo: Toward hierarchical content organization for efficient browsing, summarization and retrieval, IEEE Trans. Multimedia 7 (4) (2005) 648–666.

[15] S. Lee, M. Hayes, Real-time camera motion classification for content-based indexing and retrieval using templates, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'02), Orlando, Florida, 2002, pp. 3664–3667.

[16] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, Mean shift based nonparametric motion characterization, in: Proceedings of International Conference on Image Processing(ICIP'04), Vol. 3, Singapore, 2004, pp. 1597– 1600.

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

[18] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (2) (1994) 191–234.