Statistical motion-based object indexing using optic flow field

R. Fablet[1] P. Bouthemy[2]
[1]IRISA/CNRS [2]IRISA/INRIA
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
e-mail: {rfablet,bouthemy}@irisa.fr

Abstract

In this paper, we propose an original approach for content-based video indexing and retrieval. It relies on the tracking of entities of interest and the analysis of their apparent motion. To characterize the dynamic information attached to these objects, we consider a probabilistic modeling of the spatio-temporal distribution of the optic flow field computed within the tracked area after canceling the estimated dominant motion due to camera movement. This leads to a general statistical framework for motion-based video classification and retrieval. We have obtained promising results on a set of various real image sequences.

1 Introduction and problem statement

In order to cope with the increasing development of digital video libraries, new methods are to be defined to access and manipulate this tremendous amount of information, which implies a content-based analysis of these visual documents [1]. The first step consists in extracting the elementary temporal units (shots) which compose the video. Then, the content of each extracted shot is characterized based on key-frame selection [4], mosaic image construction [7], extraction and tracking of entities of interest, [2, 5]. The description of the motion content attached to entities of interest usually consists in determining the trajectory of these objects exploiting 2D parametric motion models and in extracting qualitative pertinent features such as the direction of the displacement [5]. However, in case of complex motion (fluid motion, crowds, sport events), motion information cannot be easily handled in such a way. Therefore, we aim at deteriminning a new characterization of the motion distribution attached to the considered entities of interest. We compute the residual optic flow field within the tracked area after canceling the estimated dominant motion due to camera movement. Exploiting a statistical framework, we consider a hierarchical motion-based classification stage, and we define a statistical retrieval scheme with query by example.

This paper is organized as follows. Section 2 describes how we interactively extract entities of interest and automatically track them in a video shot. In Section 3, the statistical modeling of the motion distribution based on the estimation of residual optic flow fields is presented. We introduce the statistical framework for motion-based classification and retrieval in Section 4. Finally, Section 5 contains experimental results and concluding remarks.

2 Tracking entities of interest

The first step of our indexing scheme consists in extracting entities of interest in the video shots. To this end, we consider the semi-automatic tracking technique presented in [6]. A region is specified by the user through an interactive interface by pointing the vertices of the bounding polygon. Then, at each instant, the dominant motion, represented by a 2D affine motion model, is estimated over this region using a robust estimator (see subsection 3.1) and used to project the polygon in the next image which constitutes the new position of the tracked entity and the new support to compute again the dominant motion at the next instant.

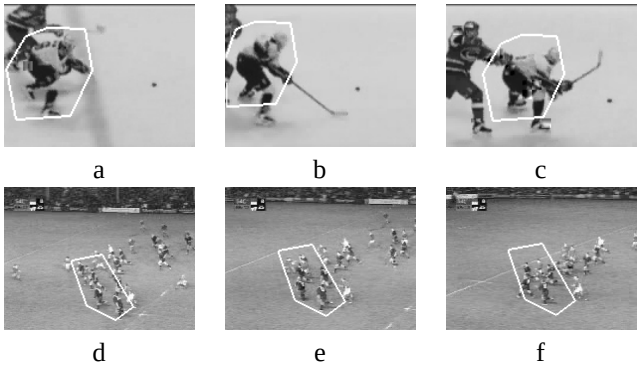

Figure 1. *Results of tracking of a hockey player (a-b-c) and of a specified area of a rugby playing field (d-e-f). The bounding polygon encompassing the tracked area is displayed in white.*

This tracking technique can cope with a variety of challenging situations such as complex motions, changes in illumination or partial occlusions [6]. In Fig. 1, we display two examples of tracking. The first one involves a hockey player and the second one a specific area of a rugby playing field in a video shot acquired with a mobile camera. Whereas the first case could also be processed using a motion de-

tection module even if it represents a not so easy situation, the second one cannot be addressed with usual techniques developed for motion segmentation. This tracking module offers new functionalities in the context of video indexing, especially considering the formulation of video queries. In many situations such as sport events, the information of interest can be not only a single player but also a group of players or more generally a given zone of the scene.

3 Motion distribution characterization

We aim at adapting statistical tools presented in [3] to cope with motion characterization within a tracked area in the video shot. In [3], our approach relies on the analysis of the spatio-temporal distribution of local motion-related measurements directly computed from the spatio-temporal derivatives of the image intensity function. When focusing on an area of interest, the use of dense optic flow field becomes reasonable w.r.t. required computational load, while providing complete motion information in terms of direction and magnitude. Besides, we want to evaluate the actual motion of the tracked object. To this end, we compute the residual optic flow field after canceling the estimated dominant image motion assumed to be due to camera movement.

3.1 Dominant motion estimation

We model the dominant motion between two successive images as a 2D affine motion model. The estimation of the six affine motion parameters is achieved with the robust gradient-based estimation method described in [10]. The use of a robust estimator ensures the motion estimation not to be sensitive to secondary motions due to mobile objects in the scene. The minimization is performed by means of an iterative reweighted least-square technique embedded in a multiresolution framework.

3.2 Residual optic flow estimation

To estimate the residual optic flow field, we exploit the technique described in [8]. The problem is stated as the global minimization of an energy function which involves robust estimators both in the regularization energy term to preserve motion discontinuities and in the data-driven energy term to discard the optic flow constraint when not valid. This minimization is efficiently performed through a multigrid algorithm which exploits different levels of local parameterization of the flow field.

In practice, when considering two successive images of a video, we first compute the dominant image motion. Then, focusing on the area of interest, the flow field resulting from the estimated 2D affine motion model is used as an initialization for the dense optic flow field estimation in this area. Finally, the residual optic flow field to be used in the motion characterization stage is the difference between the computed dense optic flow field and the dominant motion field.

3.3 Statistical motion distribution modeling

The characterization of motion information within the area of interest relies on a statistical modeling of the distribution of the occurrences of the computed residual optic flow field. Such a statistical representation will be denoted as model V. The conditional likelihood $P_V(o)$ of sequence of residual optic flow fields o computed within the tracked area over the shot is given by:

$$P_V(o) = \prod_{k=1}^{k=K} \prod_{r \in \mathcal{R}_k} P_V\left(o_k(r)\right) \tag{1}$$

where K is the length of the shot, $\mathcal{R}_k$ the tracked region in image k of the shot, $o_k(r)$ the residual velocity at point r in image k and $P_V\left(o_k(r)\right)$ the conditional likelihood of the occurrence of velocity $o_k(r)$ w.r.t model V.

To estimate model $\widehat{V}$ for a given tracked area over a shot, we exploit a Parzen window density estimator. Since we consider a simple kernel estimate, it comes to compute the histogram H^o of the occurrences of a quantized version of residual optic flow fields o within the tracked area. Let us note Λ the range of quantized residual velocities (in practice, the horizontal and vertical components of the residual optic flow vectors are quantized over sixteen levels within $[-8, 8]$). Model estimate $\widehat{V}$ is then defined by the set of model components $\left\{P_{\widehat{V}}(\lambda)\right\}_{\lambda \in \Lambda}$ given by:

$$\forall \lambda \in \Lambda,\ \ P_{\widehat{V}}(\lambda) = H^o(\lambda) / \left[\sum_{\lambda' \in \Lambda} H^o(\lambda')\right] \tag{2}$$

In addition, given a model V and a sequence of residual optic flow fields o, the conditional log-likelihood, $LF_V(o) = \ln P_V(o)$ issued from relation (1), can be simply expressed as a dot product between histogram H^o and the set of model potentials $V\ \{\ln P_V(\lambda)\}_{\lambda \in \Lambda}$:

$$LF_V(o) = \sum_{\lambda \in \Lambda} H^o(\lambda) \cdot \ln P_V(\lambda) \tag{3}$$

4 Statistical motion-based object classification and retrieval

The statistical representation of the spatio-temporal motion distribution within the tracked region can be exploited for motion-based video indexing and retrieval using partial query. Considering a set of video sequences, the associated stored set of extracted regions of interest and their associated motion distributions, we aim at retrieving, from this

database, examples similar to a video query in terms of motion content. The general idea is to define an appropriate similarity measure between shots and to determine the closest matches according to this measure.

As far as query formulation is concerned, we can also supply the user with an interactive interface to express the video query. Using the tracking module presented in Section 2, the user can specify an area of interest in the first image of the query shot and this region is automatically tracked along the shot. Then, the residual optic flow fields within the tracked region are computed at the successive instants of the shot. We will consider this tracked area as the video query in the sequel. This scheme allows the user to formulate a wide range of partial video queries such as tracking specific entities (objects, characters) or focusing on a particular area of the scene involving different entities.

4.1 Bayesian retrieval

Similarly to [3], we formulate the retrieval process as a Bayesian inference issue. In fact, considering a query q and a stored set of videos $(n)_{n\in\mathcal{N}}$, we determine the best matches n^* according to a MAP criterion expressed using the Bayes rule as follows: $n^* = \arg\max_n P(q|n)P(n)$

The distribution $P(n)$ represents the a priori knowledge on the processed database. In our case, we will introduce no a priori, which implies an uniform distribution $P(n)$. Knowing the statistical model V^n attached to an element n of the video base and computed as described in Section 3, $P(q|n)$ is expressed as the conditional likelihood $P_{V^n}(o_q)$ of the quantized residual optic flow fields o_q estimated for query q, *w.r.t.* model V^n:

$$n^* = \arg\max_n P_{V^n}(o_q) \tag{4}$$

4.2 Hierarchical classification and retrieval

To handle large video databases, it is generally required to build a hierarchical structure of the database related to content similarity. To this end, we consider an ascendant hierarchical classification scheme [3]. It consists in iteratively forming new clusters in the hierarchy of entities of interest within video shots of the processed base by merging in turn pairs of elements which minimize a given similarity measure. This process results in a binary tree.

Considering two elements of the database n_1 and n_2, their associated models V_{n_1} and V_{n_2} and the quantized residual optic flow fields o_{n_1} and o_{n_2}, the similarity measure $D(n_1, n_2)$ is a symmetric version of the Kullback-Leibler divergence $KL(n_1\|n_2)$:

$$D(n_1, n_2) = \frac{1}{2}\left(KL(n_1\|n_2) + KL(n_2\|n_1)\right) \tag{5}$$

$KL(n_1\|n_2)$ is approximated using an empirical average of the ratios of log-likelihood for the distributions attached to models V_{n_1} and V_{n_2} (see [3] for details). It indeed comes to approximate $KL(n_1\|n_2)$ by the following ratio:

$$KL(n_1\|n_2) \approx \ln\left(P_{V_{n_1}}(o_{n_1})/P_{V_{n_2}}(o_{n_1})\right) \tag{6}$$

When evaluating the similarity measure between two clusters C^1 and C^2, we exploit the following definition of D:

$$D(C^1, C^2) = \max_{(n_1,n_2)\in C^1\times C^2} D(n_1, n_2) \tag{7}$$

In the ascendant classification procedure, the creation of a new cluster in the hierarchy requires to attach to the corresponding node a statistical model in order to perform later the retrieval process through this indexing structure. Since the occurrence histogram of the quantized residual optic flow vectors for the regions of interest of a group of video shots is the merge of individual occurrence histograms, the associated model can be easily estimated using relation (2).

When performing retrieval operations over a given video base, its hierarchical representation is efficiently used to satisfy a video query. First, the node of the highest level of the tree which verifies the MAP criterion is chosen. Then, at each step, we select the child node of the current selected node, still using the MAP criterion, until a given number of answers or a given precision is reached.

5 Results and concluding remarks

We have carried out experiments on a set of real image sequences. We have paid a particular attention to choose a video set representative of various motion situations : sport videos (basket, rugby, hockey,...), rigid motion situations (cars, train, ...), and low motion activity examples. Finally, we consider a database of 50 video sequences.

For each element of this database, we have selected entities of interest over which we have performed the estimation of their residual optic flow fields. Then, we have computed the associated occurrence histograms and ML statistical models. Afterwards, we have applied the hierarchical motion-based classification scheme. In Fig. 2, we report three examples of retrieval operations. For each query, four answers are sought. The first example involves a news program where the entity of interest is the anchor. It presents a very weak motion activity. The four retrieved answers belong also to this class of video. The second example is a shot of a hockey game with a close-up on a player. The proposed answers involve also video with important motion activity and a focus on a particular player. The last query is concerned with a specific part of the playing field in a hockey game. The system again delivers correct examples similar to the query in terms of motion content, since they also correspond to a wide-angle shot of the playing field.

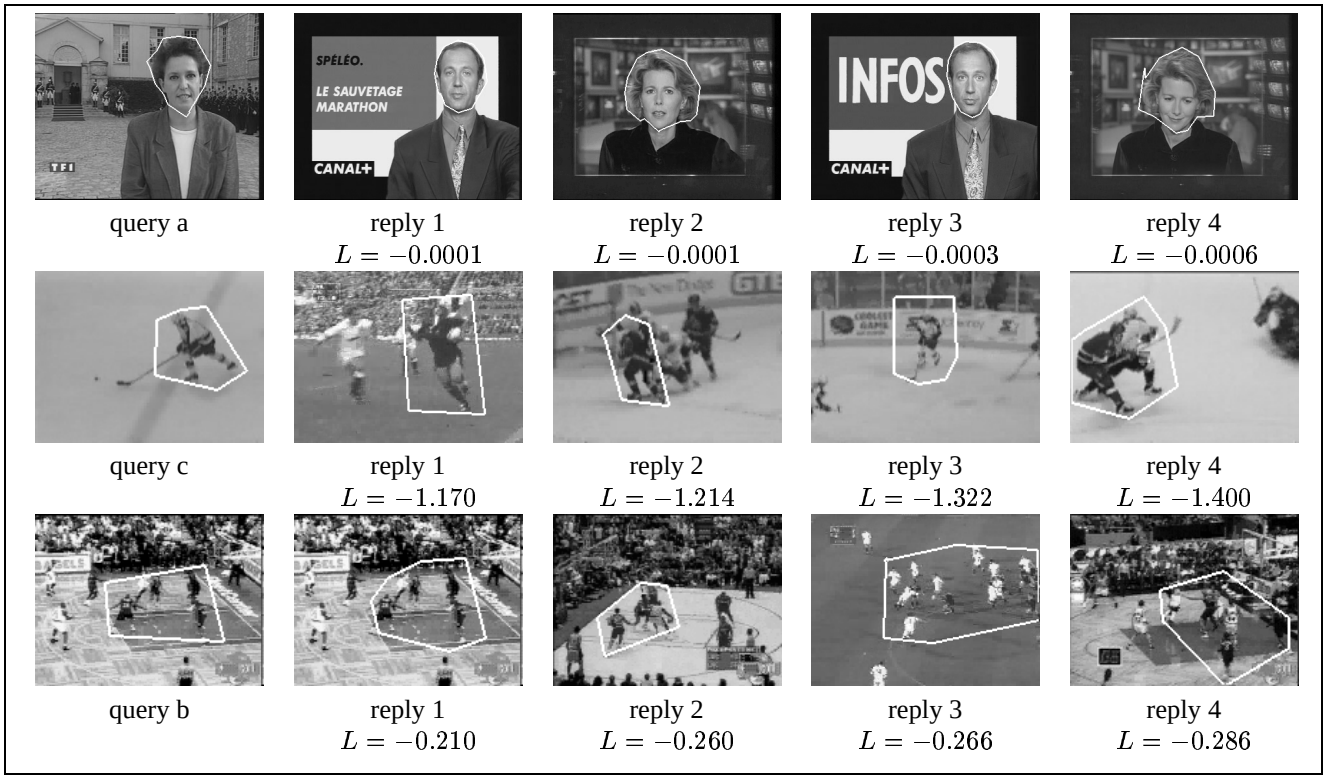

Figure 2. *Examples of motion-based retrieval operations. For each query shot and for each retrieved shot, we display the first image, and the entity of interest is delimited by a white polygon. Besides, for each reply, we report the value of log-likelihood L of the query w.r.t. statistical motion model V attached to this reply.*

We have described in this paper an original approach for motion-based video indexing and retrieval. It relies on the statistical analysis of the motion distribution of the residual optic flow fields computed within a tracked region of interest in the video shot. Exploiting a probabilistic modeling, we have established a general statistical framework for hierarchical video object classification and retrieval with query by example based on motion content. We have obtained promising results on a set of real videos. In future work, we will evaluate this approach on a larger database. We are also investigating other means of designating the region of interest involving automatic motion detection module.

Acknowledgments: The authors would like to thank E. Mémin for providing the code for optic flow estimation, and INA (Institut National de l'Audiovisuel) for supplying the MPEG-1 news sequences.

References

[1] P. Aigrain, H-J. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications,* 3(3):179–202, September 1996.

[2] J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition,* 30(4):607–625, April 1997.

[3] R. Fablet, P. Bouthemy, and P. Pérez. Statistical motion-based video indexing and retrieval. In *Proc. 6th Conf. on Content-Based Multimedia Information Access, RIAO'2000,* pp 602–619, Paris, April 2000.

[4] A. Muffit Ferman and A. Murat Tekalp. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Jal of Visual Communication and Image Representation,* 9(4):336–351, December 1998.

[5] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. 5th Eur. Conf. on Computer Vision, ECCV'98,* Freiburg, June 1998, LNCS Vol 1406, pp 595–609, Springer.

[6] M. Gelgon, P. Bouthemy, and T. Dubois. A region tracking technique with failure detection for an interactive video indexing environment. In *Proc. 3rd Int. Conf. on Visual Information Systems, VISUAL'99,* Amsterdam, June 1999. LNCS Vol 1614, pp 261–268, Springer.

[7] M. Irani and P. Anandan. Video indexing based on mosaic representation. *Proc. of the IEEE,* 86(5):905–921, May 1998.

[8] E. Mémin and P. Pérez. A multigrid approach for hierarchical motion estimation. In *Proc. 6th IEEE Int. Conf. on Computer Vision, ICCV'98,* pages 933–938, Bombay, January 1998.

[9] R. Mohan. Video sequence matching. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'98,* pp 3697–3700, Seattle, May 1998.

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