

Places Clustering of Full-Length Film Key-Frames Using Latent Aspect Modeling Over SIFT Matches

Maguelonne Héritier, Langis Gagnon, *Member, IEEE*, and Samuel Foucher, *Member, IEEE*

Abstract—An improved unsupervised classification method to extract and link places features and cluster recurrent physical locations (key-places) within a movie is presented. Our approach finds links between key frames of a common key-place based on the use of a probabilistic latent space model over the possible local matches between the key frames image set. This allows the extraction of significant groups of local matching descriptors that may represent characteristic elements of a key-place. An exhaustive evaluation of our approach was conducted on in-house and public image datasets, as well as on full-length movies. Results revealed that our method is very efficient for near-duplicate object/background detection with weak overlap. Performance measurements on full-length movies indicate a recognition rate of about 75% on the key-places clustering with a false alarm rate (FAR) of approximately 2%.

Index Terms—Duplicate detection, scene categorization, scene matching, video description, video indexing.

I. INTRODUCTION

THE INTENT OF this paper is to describe an improved approach to detect and cluster “places” in full-length movies using latent space modeling. In most movies, the storyline takes place in a set of recurrent places called key-places. Key-places carry important high-level semantic information that is useful for video indexing/retrieval applications and locating structures in video data [1], [6], [13], [28], [33], [40]. Key-places identification is also a very important element in computer-assisted production of video description (also known as described video, video narration, and audiovision) [15], [16]. Video description entails adding narration to the audio track which verbally describes visual elements of the multimedia for the blind and visually impaired people. This industry is growing owing to the imposition of regulations to increase broadcasting of programs with video description. This paper presents a component of a larger project that targets the development of software tools for computer-assisted video description [16].

Changes in camera viewpoints, foreshortening, scaling, occlusions, lighting, etc., increase the complexity of automatic place recognition in movies. Place identification is extracted

Manuscript received September 13, 2007; revised May 23, 2008. First version published March 16, 2009; current version published June 19, 2009. This paper is supported in part by Department of Canadian Heritage through the Canadian Culture Online program and Ministère du Développement Économique de l’Innovation et de l’Exportation of the Gouvernement du Québec. This paper was recommended by Associate Editor Y. Rui.

The authors are with the Research and Development Department, the Computer Research Institute of Montréal, Montréal, QC, Canada, H3A 1B9 (e-mail: maguelonne.heritier@crim.ca; langis.gagnon@crim.ca; samuel.foucher@crim.ca).

Digital Object Identifier 10.1109/TCSVT.2009.2017304

from the image background, which is often out of focus and usually provides only a partial representation of the location. There is often very little visual information common between same-location shots which can be used for detection. The task is complicated further since no prior knowledge about the location is available. As a result, typical object-recognition methods cannot be used directly.

A. Related Works

There is minimal research currently available specifically targeting segmentation and classification of film shots in terms of similar physical locations. One preliminary paper worth mentioning is by Aner and Kender [1] (although an earlier example is [13]), where image mosaics are built for panning cameras and matched using color histograms for spatial blocks. This method requires sufficient overlapped portions between mosaics in order to make a match. The minimum required overlap size is two thirds of the original frame size. This ideal case does not often appear in real movies.

Other approaches endeavor to find semantically connected consecutive blocks of shots termed *logical story units* (LSU) or scenes [36], [40]. An LSU is defined as a series of shots that communicate a unified action with a common locale and time. LSU segmentation is mainly based on global visual and temporal shot description. Audio-cut detection has also been used [21]. Although, LSU techniques could help to group consecutive shots from the same location, these techniques are not adaptable to our task because recurrent location can appear in different LSUs within a movie. Furthermore, LSU segmentation is still an open issue that key-place detection can help to resolve.

Place recognition has mostly been studied in the mobile robotics community, where the problem is known as “topological localization.” Much of the research has concentrated on image sequences. Image sequence sets in robotics have a natural (spatio-temporal) continuity and ordering that allow the use of baseline algorithms between the consecutive frames. Thus, many successful systems use this ordering as a means of “sewing” together long feature tracks, and generating initial structure and camera estimates. Given these estimates, further matches may then be sought between non-contiguous frames based on an approximate spatial overlap of the views. In full-length movies, we have discontinued sets of ordered frames that are relative. In addition, we cannot know *a priori* whether the between-shot view motions are small.

Our task is mainly related to recent works regarding near-duplicate detection [20], [28]. More specifically, it is related

to the near-duplicate object-detection tasks or near-duplicate background-detection tasks. This type of near-duplicate detection is composed of footages or images of the same object or same background but taken at different times and/or different places. Matching the images through on interest- or key-points methods are often used because they are robust to occlusions and illumination changes. Moreover, invariant descriptors for the key-points also make these methods robust to view point change. There are mainly two different groups of approaches based on key-point matching techniques which have been proposed in the literature. One group (e.g., [3], [7], [8], [9], [18], [20], [30], [35], and [42]) consists of filtering out the outliers using robust matching methods, such as random sample consensus (RANSAC) [14] or least median of squares [27], between the entire image. However, those fitting methods perform poorly when the ratio of inliers falls below 50%. This means that a large overlap between pair of images is required for the matching—which is rarely the case for location representation in a movie. In [22], Lowe proposed to cluster features in the pose space using the Hough transform. This method requires the setting of many parameters which limit robustness. The second more recent group of approaches tries to find common spatial patterns (e.g., [29], [30], and [32]–[34]). Those approaches are mainly based on comparing key-point neighborhoods. However, there is an ambiguity in the choice of the neighborhood size used for the comparison. Moreover, outliers can be present in the neighborhood. In fact, there will always be wrong matches possible based on very common local structures. To resolve this issue, some authors [33], [34] use an efficient representation, inspired from text analysis, to represent neighborhoods called “Bag Of Words” (BOW). BOW consists of representing a text document as a vector, counting the number of occurrences of different words as features. In [33] and [34], descriptors are quantized into clusters which are analogous to “words” in a text document.

The BOW representation has two drawbacks when dealing with ambiguities: *polysemy* (i.e., a word that has two different meaning); and *synonymy* (i.e., two words with the same meaning). BOW generative models capture the co-occurrence information between elements in a collection of discrete data by introducing a latent variable (i.e., a context value) in order to raise the ambiguities of the BOW representation. Probabilistic latent semantic analysis (pLSA) is one of the best-known BOW models [17]. BOW generative models are used in natural language processing and statistical text analysis to discover topics in documents [17]. They have recently been applied to classification in the image processing field [6], [10], [11], [12], [26], [31]. Local patches are called *visterms* and are modeled as basic building blocks of an image, which are analogous to words in text documents. Images are represented as a collection of *visterms*. BOW models are then used to extract distributions of characteristic *visterms* which are denoted as topics (e.g., grass, building, wheel), that are shared across images. Topic representation of an image is used to perform scene or object classification.

B. Overview of Our Method and Main Contributions

We want to resolve near-duplicate objects/backgrounds detection when the overlap between images is weak. This kind

of detection in a movie is important because recurrent locations appear throughout many different view points. There is often very little visual information shared between key-frames of a given location. Thus, inliers are often hidden among a large number of outliers. Existing methods cannot efficiently handle near-duplicate background detection under these conditions.

Our approach is based on a BOW representation of sub-images. We use a BOW generative model to filter out wrong matches, based on very common structures, and extract groups of matches that may represent a key-place semantic characteristic. We extract inliers, at whatever the outlier rate may be, by introducing a latent value for each match. A latent value is a context value shared by a group of local matches that may represent a key-place semantic characteristic (analogous to a “topic” for text document). We use the latent Dirichlet allocation (LDA) generative model [5], which is a new generative model derived from pLSA [17]. LDA has been shown to be superior to pLSA because it can be applied to documents that contain several topics and can generate documents that are not in the training corpus. We use LDA to extract significant matches’ distribution between key-images (key-frames) of a film. This generative model provides a discrete discriminant analysis over matches. The *visterms* are seen as a group of local descriptors that match. Significant *visterm* distributions extracted are seen as a part of latent “topics,” which are, in fact, typical representative elements of a scene in a higher semantic level (e.g., cigarette shelves, coffee machine). Latent topics are used as context values for *visterms*. A group of local matches (*visterms*) sharing the same latent topic constitutes a “topic link” across images. Fig. 1 illustrates the concept of latent semantic analysis and gives a block diagram summarizing our approach.

LDA was previously applied by Sivic *et al.* [31] for unsupervised object classification to find discriminant word distribution for each object category. Here, we use LDA to extract significant group of correspondences and filter out wrong correspondences. In full-length films, a place cannot be described by one topic, as it is often represented under many different view points. For the application at the origin of our work (development of a software tool to computer assist post-producers in the production of video description for blind people), we often have to deal with only one reoccurrence of one weak visual characteristic to assign a key-frame to a key-place. The main contribution of our paper is an improved way of combining LDA and scale-invariant feature transform (SIFT) to 1) get a recall improvement of near-duplicate objects or background detection when the overlap between images is weak and 2) filtering out the many SIFT outliers resulting from our choice to privilege a high correspondence-detection rate in order to get the maximum of correspondences in key-frames with small overlap.

The paper is organized as follows. Sections II and III give details about the methodology and implementation of our approach. Section IV presents the evaluation process and performance results on in-house and public image datasets, as well as on full-length movies. Finally, Section V discusses the possible avenues for future works to address performance and other improvements.

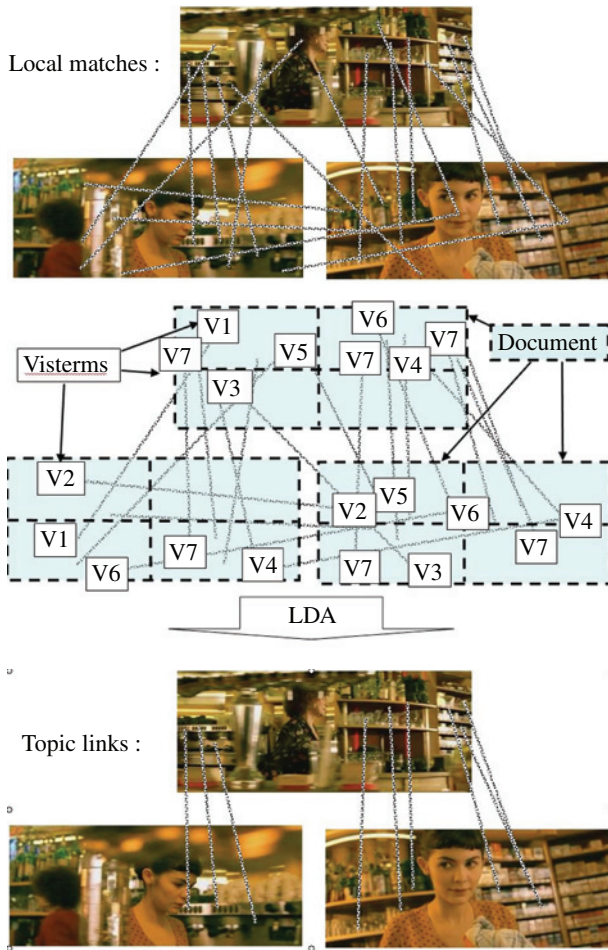


Fig. 1. Principle of latent semantic analysis and a block diagram that summarizes our approach.

II. METHODOLOGY

Our key-place software module is made of four sub-modules (see Fig. 2). First, video shots are summarized into a few relevant frames. Second, a coarse link extraction between shots is performed using global visual images features and temporal information. Third, fine links are extracted on the basis of finding significant groups of local descriptors' matches using latent aspects. Finally, a hierarchical clustering of shots belonging to the same key-place is done from the links extracted from the previous step. The module inputs are the shot transitions and locations of actor's face regions—which are both detected automatically. The key-place module avoids processing face regions in order to preclude the LDA from the

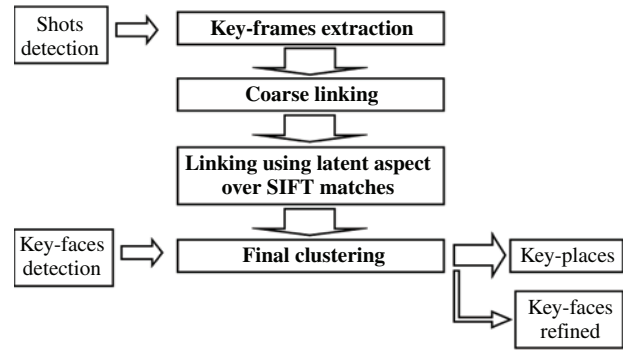


Fig. 2. Block diagram of the key-place software module.

assigning faces as a topic link for key-places, thus limiting false alarms.

A. Key-Frames Extraction

After an automatic shot transition detection step, each shot is summarized in few representative frames (key-frames). To this aim, we compute the overlap between images using a simple method based on camera motion estimation [25]. The algorithm finds the optimal frame path over the shot, which minimizes the overlap between frames. Small shots that are less than 10 frames in size are discarded.

B. Coarse Linking Using Low-Level Image Descriptors

This is a preprocessing step for the LDA-based link extraction. The goal is to extract trivial links between similar key-frames that do not require the application of the more complex LDA linking algorithm. For example, there are often similar key-frames during bits of dialog where the camera goes back and forth from one person to another. These key-frames bring no new visual content. In these cases, we use color and edge histogram distances and restrict the linking process to a limited temporal distance around the considered key-frame. These features are sufficient to link very similar frames.

There are two advantages in this preprocessing step. First, it saves computation time by reducing the number of frames to process for the further more complex steps. Second, it forces the LDA to concentrate on nontrivial links (LDA converges easily with trivial links).

C. Key-Frames Linking Using Latent Aspects on SIFT matches

This is the main step of our algorithm which extracts groups of matches between key-frames. We use the concept of “Bag of visterms” (BOV) to represent each key-frame, in conjunction with a different method of building the representations based on the K-nearest neighbor (K-NN). We then apply a generative probabilistic model to extract groups of local matches that represent a characteristic structure of a specific key-place.

1) *Image Set Representation*: The construction of BOV is done from a set of several key-frames extracted from each shot. First, regions of interest (ROIs) are automatically detected in

the image with a difference of Gaussians (DOG) point detector over which one computes local descriptors using SIFT [22]. We use SIFT because it performs the best in terms of specificity of region representation and robustness to image transformations [24]. We tried several other descriptors (maximally stable extremal regions (MSER) based [23], PCA-SIFT [20], etc.) in a preliminary work [15] but found that SIFT with DOG was the best to establish the difficult correspondences between images of various appearances and quality (low illumination, night, smoothing filter, etc.). For instance, only the SIFT with DOG algorithm was able to detect correspondences between the first two images in Fig. 5. However, we had to choose a very weak selection step on SIFT features' matches in order to obtain difficult but important correspondences. It became apparent that high-quality correspondences are hidden among a large number of outliers. We configured the algorithm to privilege a high correspondence rate over outliers in order to acquire the maximum number of correspondences within key-frames with the smallest overlap. Finding an efficient method to filter out the outliers is one of the main contributions to this paper.

Second, in order to obtain a text-like representation, descriptors must be clustered. We do not use centroid clustering such as K-means for descriptor quantization [6], [12], [26], [32], [33], [34]. This kind of quantization consists of identifying several clusters within a training set. Each descriptor is then assigned to the nearest cluster and generates matches in conjunction with all descriptors assigned to this cluster. This quantization approach is fast but introduces errors, since cluster centroids may be not well defined. Also, the cluster centroid values are only a coarse representation of the clustered descriptors. As an alternative to reduce errors, we use K-NN between SIFT descriptors belonging to different images to create the visterms. Therefore, a visterm is a set of matched local descriptors from different images. The K-NN is used to match normalized SIFT descriptors between two different images based on the Manhattan distance. K-NN is computationally expensive. We use an approximate K-NN approach based on a priority tree search [2].

Bad matches are discarded when their distances are above 0.6, and when the distance to the first NN is above 90% of the distance to the second NN. Finally, a BOV representation h is built from the local descriptors according to

$$h(d) = \{h_i(d)\}_{i=1..N_V}, \quad \text{with } h_i(d) = n(d, v_i) \quad (1)$$

where $n(d, v_i)$ is the number of visterm occurrences v_i in a sub-image d , and N_V is the size of the vocabulary V (i.e., the set of all visterms). This representation contains no information about the spatial relationships between visterms, the same way the standard BOW text representation removes the word ordering information. This is the reason why we use a sub-image instead of the entire image. Each image is divided into several sub-images of different sizes (with or without overlapping). Sub-image sizes are chosen to capture specific visual characteristics that could refer to a location. This partition is based on what is visually identifiable by a human.

2) *Generative Model*: An excellent presentation of the pLSA and LDA models applied to visual content can be found

in [31]; here, we only give a brief overview of the main concepts.

According to the pLSA [17] framework, we have sub-images represented as documents, and we want to discover topics as semantic characteristics of location (e.g., shelves, tapestry from a room, etc.) such that two images sharing instances of semantic characteristics are modeled with shared topics. Those shared topics form a topic link. The models are extracted from the BOV representation of sub-images. Visterm analogs of a word are formed by SIFT-matching feature descriptors. Let a collection (corpus) of sub-images $D = d_1, \dots, d_{N_D}$ with visterms from a visual vocabulary be $V = w_1, \dots, w_{N_V}$. One can summarize the data in a $N_D \times N_V$ co-occurrence table of counts $h_i(d_j) = n(d_j, w_i)$. The pLSA [17] model establishes that a document labeled d and a visterm w_n are conditionally independent, given an unobserved topic z . A joint probability model $P(w, d)$ over $N_D \times N_V$ is defined by the mixture

$$P(w, d) = P(d) \sum_{z \in Z} P(z|d) P(w|z) \quad (2)$$

where $P(w|z)$ are the specific topic distributions. Each image is modeled as a mixture of topics $P(z|d)$ [17].

In pLSA, no document, which is represented as a list of numbers (the mixing proportions for topics), was considered in the generative probabilistic model. This leads to several problems: 1) The number of parameters in the model grows linearly with the size of the corpus, which leads to serious over-fitting; and 2) it is not clear how to assign probability to a document outside of the training set.

LDA is a corpus-generative probabilistic model [5]. LDA allows each document to exhibit multiple topics with different proportions and, therefore, can capture the heterogeneity in grouped data that exhibit multiple latent patterns. The basic idea is that documents are represented as random mixtures over k latent topics, where each topic is characterized by a distribution over words. The framework treats the topic mixture weights as a k -parameter hidden random variable (θ) and places a Dirichlet prior (α) on the multinomial mixing weights. The word probabilities per topic are parameterized by a $k \times N_V$ matrix β . The model parameters (α and β) are estimated using the maximum likelihood principle over a set of training sub-images D . Optimization is performed using a variational expectation-maximization (EM) algorithm. By using an approximation inference algorithm, these independent sub-image parameters can then be used to infer the document-level parameters (related to θ and z) of any sub-image, given its BOV representation $h(d)$.

D. Final Hierarchical Clustering

Grouping of the key-frames belonging to a same key-place is then performed by constructing a graph between shots using links extracted by the method previously described. Clusters (sub-graphs) are identified using a bottom-up hierarchical clustering algorithm, which consists of assigning a key-frame to a key-place cluster if the frame has a number of connections to other images within the key-place cluster above a given threshold.

III. IMPLEMENTATION CONSIDERATIONS

The larger the set of images, the more connections are possible between the SIFT descriptors and, therefore, between the visterms. Therefore, the number of visterms tends to converge to one when the size of the image set increases, which is a consequence of the decrease of groups of descriptors that do not match. Thus, in order to keep a sufficiently large vocabulary size and avoid all descriptors being quantized to a unique visterm, we had to limit the size of the image set. We then divided the initial image set into several random smaller sets of images (about seven or eight images in each of our tests).

We also used different sizes of sub-images in the key-frames, which were constructed from different grid decompositions and various link extraction resolution levels. A small sub-image size and a high key-frame resolution allow extracting links with few matches on very specific image details. Conversely, a larger sub-image size allows extracting more global image characteristics. Experience has taught us that results are better when different sub-images sizes are used for the LDA modeling step. Thus, we set the sub-image size to one-third of the original image size, and one-fourth the size in the case of low-resolution images. Sub-images size for the high resolution levels were chosen at one-fifth and one-eighth the image size.

The LDA algorithm requires setting the number of topics. This is set automatically during the initialization step. Topics are initialized by assigning visterms from a random document to a topic until no more visterms are available. An external parameter controls the topic-fragmentation level during the initialization step by setting a threshold value (0.3 in our tests) for the rate of word overlap between two different topics. The LDA model parameter α is initialized to 0.1.

Before running the LDA learning process, key-frames are divided into sub-images. Each sub-image is described by its BOV. After application of the LDA, we select the best sub-images and visterms for each topic. Overlapping sub-images sharing more than two common visterms from the same topic are merged. Then, a topic link is formed when two of the selected sub-images share more than four selected visterms. A further step is added to filter wrong links. It consists of eliminating topic links for which SIFT matches are not within the same range of scale variation.

IV. TESTS AND EVALUATION

This section presents linking and clustering performances of our method on two dataset types.

- 1) An in-house dataset consisting of 11 key-places classes (five outdoor, five indoor, and a “noise” class). Each indoor and outdoor class is made of 10 images with different viewpoints and illuminations. The noise class is composed of 10 random indoor images and 10 random outdoor images (see Fig. 3). Images are 640×480 pixels in size.
- 2) Two full-length French movies of about 1.75 h each: “Le fabuleux destin d’Amélie Poulain” (film 1) and “La vie est un long fleuve tranquille” (film 2). Key-frame size is 700×350 pixels.



Fig. 3. Samples of our in-house place image dataset. The set is publicly available at <http://www.crim.ca/vision>.

We also compared our performance against other approaches on the above datasets as well as two additional public image datasets of places, namely the “Raglan” and the “Valbonne” datasets [30]. Before discussing results, we propose a new measure to quantify the *a priori* complexity of a dataset.

A. Complexity Measures

Performance measures can differ significantly for different datasets. Overlap between images, illumination changes, occlusions, and point-of-view variations influence the complexity of the linking and clustering. We therefore found it important to derive a linking complexity measure (LCM) and a clustering complexity measure (CCM) of a dataset.

LCM is calculated over (rectangular) overlapping regions (see Fig. 4) and is the ratio between the number of point descriptors outside the extracted region and the mean number of point descriptors within the two entire images. The number of point descriptors in a region is an indicator of the amount of information present. This measure does not take into account the viewpoint and lighting variations. However, it gives a general idea regarding the difficulty in matching two images.

The CCM is the couple formed by the mean and variance of the LCM, calculated for all image pairs in the considered cluster. When the mean value is high, the cluster is expected to be difficult to build. A high variance value indicates that there are both easy links and complex links. Combined with a high mean value, this may lead to overclustering.



Fig. 4. Example of overlapping content for the calculation of the LCM.

TABLE I
LINKING PERFORMANCE MEASURE FOR OUR IN-HOUSE DATABASE

Complexity %	Total			High resolution		Low resolution	
	GT	P	R	P	R	P	R
<60	4	1	1	1	0.75	1	1
6070	6	1	0.83	1	0.5	1	0.67
7080	23	0.96	0.57	0.95	0.57	1	0.43
8090	41	0.93	0.51	0.97	0.49	0.94	0.39
> 90	136	0.94	0.25	0.96	0.21	0.96	0.1
All	210	0.94	0.37	0.96	0.33	0.96	0.23

“GT” is the number of links determined as the ground truth. “P” is the precision rate. “R” is the recall rate.

B. Linking Performance

Table I shows the linking performance on our in-house database for different linking complexity levels.

We found that full-resolution images were more suitable to find links with very few local matches constructed from very specific image details. Half-resolution images were more appropriate to find links on more global image characteristics. It was also interesting to combine link extractions found for both resolution levels (see “Total” in Table I).

C. Key-Places Clustering

We used three measures for evaluating the final key-places clustering performance on the datasets: top clustering rate (Top CR); recognition rate (RR); and false alarm rate (FAR). The Top CR is the RR of the image within the largest observed cluster for a given place. RR is the recognition rate of the image within all observed clusters for a given place.

1) *In-House Dataset*: Table II shows the clustering performance on our in-house dataset for two different resolutions. The “Total” column is the clustering measures when the high- and low-resolution link extractions are combined. We have chosen a very permissive threshold value for the final hierarchical clustering step. This means that almost no links were discarded during the process. However, we need to take care not to choose a threshold value that could force a null FAR. This would mostly change the Top CR but would not affect the RR much. The “Apartment” class gives the worst results. This

could be explained by the fact that this place does not show up consistent visual characteristics. Indeed, the specificity of this place is mainly due to the furniture arrangement. Invariant key-point features from furniture corners are not consistent between different views. Invariant feature matches are better adapted for planar visual characteristic structures such as tapestry, picture, garnishment, or object texture.

2) *Full-Length Movies Dataset*: For film 1, we automatically extracted 1223 shots and 1561 key-frames. The four main places of the film are the bar (20% of the shots), the neighbor’s apartment (11%), Amélie’s apartment (8%), and the grocery (4%). Eight hundred and twenty-two out of the 1223 shots are considered as a part of one of the 35 key-places defined in the ground truth (GT). The rate of wrong links between shots is 8% before hierarchical clustering. Thirty-three of the 35 key-places in the GT are represented in one or several clusters. Table III shows measures for the four top places in the movie. A place can be represented by several key-places in the GT (e.g., kitchen and bedroom for Amélie’s apartment). Indeed, we cannot have a visual link between the kitchen and the bedroom because these are two different rooms. Amélie’s apartment is represented by five clusters in the GT.

For film 2, 454 shots and 700 key-frames have been automatically extracted (see Table IV). The four most important places are “Quesnoy’s ground floor” (28% of the shots), “Groseille’s apartment” (21%), “Quesnoy’s garden” (5%), and “the Doctor’s office” (4%). Four hundred and eleven shots were considered as a part of one of the 35 key-places defined in the GT. The rate of wrong links between shots is 4% before hierarchical clustering. Thirty key-places in the GT are represented in one or several clusters.

The high number of clusters per class could be explained in part by the fact that the GT was not made by taking into account whether or not there is a possible visual match between images. “Amélie’s apartment,” “Neighbor’s apartment,” and “Bar” classes have a large variability because of their many viewpoints. Also, a place sometimes appears in several close-up shots. Only few interactions between actors happen in Amélie’s apartment and this place often appears in short or close-up shots. The lower value of RR for film 2 can be explained by the fact that there are several small key-places (less than 50 shots). The RR for this kind of key-place is generally low because there are fewer occasions to extract their visual characteristics. The non-retrieved key-places (two for film 1 and five for film 2) were not well represented in the films (less than 10 shots). Despite this, the RR is quite satisfying (78% for film 1 and 74% for film 2) with a low FAR (0.1% and 0.25%, respectively).

D. Comparison With Other Approaches

First, we compared the linking part of our approach against similar approaches that also try to find common spatial patterns over neighborhood correspondences with no filtering [33], [34], filtering based on geometry estimation with RANSAC [29], [30], and for various neighborhood sizes (see Table V). We simulated these approaches by first removing our filtering step implemented with the LDA modeling for the half-resolution images on our in-house dataset. Then, we replaced

TABLE II
IN-HOUSE DATASET CLUSTERING PERFORMANCE MEASURES

Places	Complexity		Total				Full resolution				Half resolution			
	Mean	Var	RR	TopCR	Nb. of Clusters	FAR	RR	TopCR	Nb. of Clusters	FAR	RR	TopCR	Nb. of Clusters	FAR
Total			0.8	0.64	19	0.05	0.76	0.53	21	0.03	0.65	0.4	21	0.03
Apartment	0.94	0.07	0.2	0.2	1	0.5	0.2	0.2	1	0	0.2	0.2	1	0.5
Office	0.92	0.1	1	0.6	3	0	0.8	0.4	3	0	0.8	0.4	3	0
Cafeteria	0.93	0.06	0.9	0.9	1	0	0.8	0.8	1	0	0.7	0.7	1	0
Church	0.95	0.07	1	0.8	2	0	0.9	0.4	3	0	1	0.6	3	0
Square	0.98	0.07	0.8	0.4	3	0	0.8	0.4	3	0	0.5	0.3	2	0
Shopping center	0.96	0.05	1	1	1	0.1	1	0.6	2	0.14	1	0.6	2	0
Mc Gill campus	0.98	0.03	0.6	0.4	2	0	0.6	0.4	2	0	0.3	0.3	1	0
Place des Arts	0.95	0.06	0.7	0.7	1	0	0.7	0.7	1	0	0.6	0.2	3	0
St Denis street	0.97	0.4	0.9	0.9	1	0	0.9	0.9	1	0	0.6	0.2	3	0
Vieux port	0.96	0.05	0.9	0.5	3	0.17	0.9	0.5	3	0.17	0.8	0.4	3	0

"# Clusters" is the number of clusters observed for a specific place.

TABLE III
PERFORMANCE MEASURE FOR "LE FABULEUX DESTIN
D'AMÉLIE POULAIN"

Places	# SHOTS	RR	# Clusters		Top CR	FAR
			in GT	Observed		
Total	822	0.78	35	131	0.47	0.01
Bar	239	0.87	1	32	0.47	0
N apart	129	0.81	1	21	0.16	0
A apart	94	0.68	5	22	0.27	0.01
Grocery	53	0.84	2	6	0.82	0

TABLE IV
PERFORMANCE MEASURE FOR "LA VIE EST UN LONG
FLEUVE TRANQUILLE"

Places	# Shots	RR	# Clusters		Top CR	FAR
			in GT	Observed		
Total	412	0.68	35	87	0.40	0.02
LG GF	117	0.70	3	24	0.15	0.02
G apart	88	0.80	1	4	0.70	0
LQ garden	21	0.76	1	8	0.20	0.02
D office	16	0.44	1	5	0.13	0

"# Shots" is the number of shots taken in a specific place. "# Cluster GT" is the number of key-places identified in GT for a specific place. "# Cluster Obs." is the number of clusters observed for a specific place.

our filtering step by one based on a geometry estimation with RANSAC. The neighborhood size is defined by the sub-image size. The LDA modeling (as a filtering step) performs better. LDA modeling discards wrong matches based on very common links and extracts characteristic correspondence patterns. This permits more independence on the neighborhood size and deals with correspondences that have larger outlier rates. Therefore, it helps to improve the recall for shots with little common visual information.

Second, Table VI compares the clustering performance (in terms of RR, Top CR, and FAR) of our approach against the same approaches above, as well as the "AutoStitch" demo [3], [7], [8], which is a no-neighborhood approach. The results are given for our three large datasets we have as well as for two public one-class datasets used in [30]: the "Raglan" and "Valbonne" datasets (note that FAR measures are not relevant

TABLE V
LINKING PERFORMANCE COMPARISON ON OUR IN-HOUSE DATASET

Sub-image size	No filtering		RANSAC		LDA	
	P	R	P	R	P	R
1/3 image size	0.12	0.26	0.12	0.21	0.25	0.26
1/4 image size	0.12	0.26	0.12	0.21	0.74	0.26
1/5 image size	0.12	0.26	0.12	0.21	0.91	0.23
1/6 image size	0.25	0.22	0.53	0.17	0.96	0.21
1/7 image size	0.86	0.18	1	0.13	1	0.17
1/8 image size	1	0.18	1	0.12	1	0.17

"P" is the precision rate; "R" is the recall rate.

TABLE VI
CLUSTERING PERFORMANCE COMPARISON ON VARIOUS DATASETS

Dataset	LDA			RANSAC		
	RR	Top CR	FAR	RR	Top CR	FAR
Raglan	1	1	n/a	0.33	0.17	n/a
Valbonne	1	1	n/a	1	1	n/a
Raglan&Valbonne	1	1	0.97	0.97	0.5	
In-house	0.65	0.4	0.03	0.44	0.27	0
Film1	0.78	0.45	0.01	0.6	0.3	0.02
Film2	0.68	0.4	0.02	0.5	0.3	0.02

Dataset	No filtering			"AutoStitch"		
	RR	Top CR	FAR	RR	Top CR	FAR
Raglan	1	1	n/a	0.89	0.26	n/a
Valbonne	1	1	n/a	1	1	n/a
Raglan&Valbonne	1	1	0.5	0.95	0.64	0
In-house	0.54	0.34	0	0.35	0.23	0.05
Film1	0.7	0.38	0.09	0.39	0.26	0.01
Film2	0, 6	0.35	0.1	0.36	0.24	0.02

for one-class and have been replaced by a "n/a" entry). The "Valbonne" dataset is composed of 15 images of a church with small view-point differences. The "Raglan" dataset has 46 images of a castle taken from various view points. For the "Valbonne" CCM, we get CCM = (0.58; 0.12) compared with an average of (0.95; 0.11) for our in-house dataset, indicating that our dataset is more complex to process. The "Raglan" and "Valbonne" classes are quite similar. They have similar local features (brick walls). Therefore, they could be easily merged. "Raglan & Valbonne" is the dataset composed of the "Valbonne" dataset and the "Raglan" dataset. Table VI shows



Fig. 5. Links for (top) “Amélie’s kitchen” class, (middle) “Bar” class, and (bottom) “Grocery” class. Matches are within the white boxes.

that our method offers the best clustering performance for all different datasets and the best tradeoff between RR and FAR, for small as well as large datasets.

We have RR, Top CR, and FAR for six different datasets and four different neighborhood filterings: LDA filtering (one-third and one-fourth image size), RANSAC [30] (one-seventh image size), no filtering [33], [34] (one-eighth image size), and “AutoStitch” [3], [7], [8] (no neighborhood). Bold highlights the best global results.

V. DISCUSSION

A. Key-Places Clustering Performance

The more often a particular location appears in the movie, the more meaningful details are presented by this key-place, thus enabling LDA to detect its content as a dominant topic. This process is similar to what humans do when they assimilate key-places. If a location is shown several times and has many details, the viewer quickly considers it as a reference location. Our approach extracts semantic characteristics, which facilitates the recognition of a specific place and can also be used to automatically describe image content details (e.g., cigarette shelves from a bar, tapestry from a room) (see Fig. 5).

We have observed that very subtle good matches can be generated between images using the 1-NN algorithm on SIFT descriptors. The LDA approach is able to separate those subtle matches from the large number of generated matches. In fact, LDA creates a link by making a discriminant analysis between trivial matches (e.g., straight lines) and those that refer to a specific descriptor structure. Fig. 5 shows examples of links for three key-places in film 1. False links usually appear in common similar visual structures, such as striated or squared structures (see Fig. 6). They are often composed of less than seven visterms.

Extracting specific visual structure works equally well for key-faces and key-objects (see Fig. 7). In particular, the

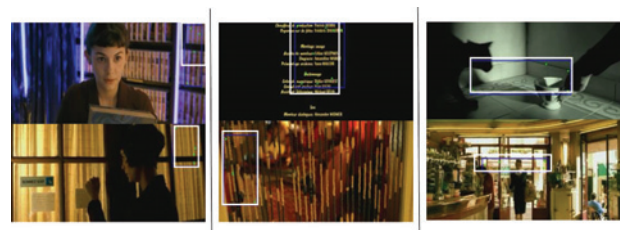


Fig. 6. Examples of wrong links.

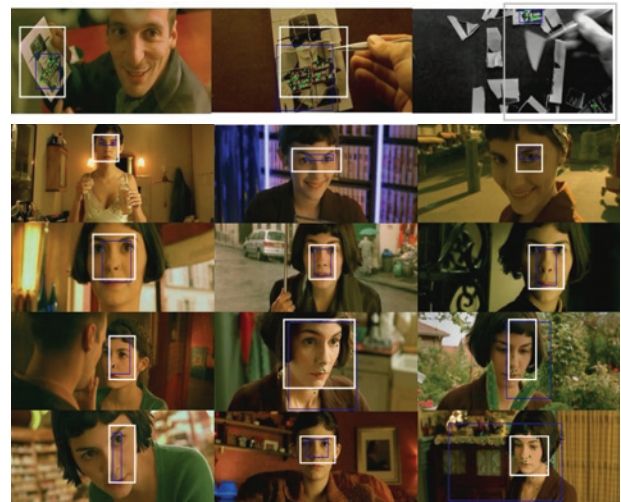


Fig. 7. (Top) Examples of potentially interesting detected key-objects; (Bottom) examples of potentially interesting detected key-faces.

method can extract faces on very close-up shots, which is a difficult task for a classical face detector. Of course, to avoid links made with faces, we first use a face detector to segment face regions, as mentioned in Section II. This opens up the possibility of using our technique for close-up face detection where classical global face detectors often fail.

Apart from identifying the actor faces as foreground, we do not attempt to identify other foreground objects at this time. It is, of course, an actual limitation of our approach, but we made this choice as we assume that it does not appear often enough to justify the additional work. Up to now, we have not faced a case where this assumption fails. Besides, our place-clustering module has been integrated into a video-summarization software where an interface allows the user to manually correct the clustering. This is a largely acceptable tradeoff from the usability point of view. However, we plan to add segmentation of moving foreground objects in the future, but segmentation of static foreground objects is still a very difficult task.

B. Computation Time

The total computation time on a dual Pentium 4 is about five times the duration of a typical full-length movie that is about 1.5 h long with 700 key-frames. This is divided as follows: one fourth of the processing time for the key-frame extraction to the SIFT calculation; one-fourth for the LDA analysis and final clustering; and one-half for the K-NN SIFT descriptor quantization. The latter is the current bottle neck of the process

and has a quadratic complexity with respect to the key-frame number. The centroid quantization approach (see Section II-C) is faster but introduces an important quantification error. However, the introduction of a latent aspect variable for each word in a document may compensate for the quantification error and thus offer good clustering performance. Centroid quantization may be interesting when processing very large video data (like rushes).

We can improve computation time (and performance) using a hierarchical approach for link extraction. The easiest links could be extracted faster at a coarser level in order to eliminate strong links that could be hiding subtle ones. More difficult links could then be extracted at a higher resolution level.

Another approach would be to use inference. Link extraction is currently performed by several LDA modeling processes. Each topic link that was correctly extracted by LDA could be used to infer similar topic links between different key-frames.

C. Improving Performance

There are several factors that could increase performance. Instead of using key-frames that could possibly discard important visual information in the shot, we could use the track representation of the features points along the shots [32], [34]. This could give a better representation of the shots and could also filter unstable features points.

We could also add color information as a feature. Indeed, color is a powerful cue in object recognition. However, the extraction and calculation of scale-invariant interest point descriptors are made on the grayscale image. Color information could be taken into account for the point descriptor [39]. In [38], Unnikrishnan and Hebert combine information from the color channels to drive the detection of scale-invariant key-points. The expressive linear model is used to compensate for the local illumination effect. This makes a descriptor robust with respect to changes in lighting without having to estimate its properties from the image.

In our method, local visterms are assumed to be independent of each other. Although this assumption simplifies computation, it does not take into account useful information encoded within the inter-relationships among the visterms. As in [41], we could introduce a linkage structure over the latent topics to encode the visterm dependency. This structure reinforces the semantic connections among the visterms by facilitating better topic clustering.

We could also take advantage of recent advances in generative modeling. One limitation of LDA is its inability to model topic correlation (for example, a document on genetics is more likely to treat the topic of disease than a document on X-ray astronomy). This limitation stems from the use of the Dirichlet distribution to model the variability among the topic proportions. The correlated topic model (CTM) allows the topic proportions to exhibit correlation via the logistic normal distribution [4]. CTM uses an alternative, which is a more flexible distribution for the topic proportions, allowing for covariance structure among the components. This gives a more realistic model of latent topic structures where the presence of one latent topic may be correlated with the presence of another.

Finally, topic distributions and visterm distributions over topics could be learned using a hierarchical structure [37].

D. Summary

We have presented an improved method to automatically cluster recurrent key-places in a movie, based on a probabilistic latent space model over local matching descriptors between the set of key-frames of the movie shots. The method is able to extract groups of significant matches that represent a semantic characteristic of a key-place. Our method seems to be more efficient than others for the near-duplicate background detection tasks with weak overlap.

ACKNOWLEDGMENT

The authors are very grateful to the reviewers for their comments on the ways to improve this paper.

REFERENCES

- [1] A. Aner and J. R. Kender, "Video summaries through mosaic-based shot and scene clustering," in *Proc. 7th Eur. Conf. Comput. Vision*, London, U.K.: Springer-Verlag, 2002, pp. 388–402.
- [2] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.
- [3] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vision*, vol. 74, no. 1, pp. 59–73, 2007.
- [4] D. Blei and J. Lafferty, "A correlated topic model of science," *Ann. Appl. Statist.*, submitted for publication.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learning Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [6] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vision*, 2006.
- [7] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vision*, submitted for publication.
- [8] M. Brown and D. G. Lowe, "Unsupervised 3-D object recognition and reconstruction in unordered datasets," in *Proc. 5th Int. Conf. 3-D Imaging Modeling*, Ottawa, ON, Canada, 2005, pp. 56–63.
- [9] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," in *Proc. Int. Conf. Comput. Vision Pattern Recognition*, San Diego, CA, Jun. 2005, pp. 510–517.
- [10] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Eur. Conf. Comput. Vision*, Prague, Czech Republic, 2004, pp. 1–22.
- [11] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proc. Workshop Generative-Model Based Vision*, Washington, D.C., 2004, pp. 178–187.
- [12] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. Int. Conf. Comput. Vision Pattern Recognition*, San Diego, CA, 2005, pp. 524–531.
- [13] A. M. Ferman, A. Krishnamachari, A. M. Tekalp, M. Abdel-Mottaleb, and R. Mehrotra, "Group-of-frame/picture color histogram descriptors for multimedia applications," in *Proc. ICIP*, Vancouver, BC, Canada, 2000, pp. 65–68.
- [14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, pp. 381–395, Jun. 1981.
- [15] S. Foucher, M. H eritier, M. Lalonde, and D. Byrns, "Literature review, design, prototyping and preliminary tests of processing algorithms for the extraction of visual content adapted to descriptive video and video mining," *Comput. Res. Inst. Montr al, QC, Canada, Tech. Rep. CRIM-06-05/11*, May 2006.
- [16] L. Gagnon, S. Foucher, F. Lalibert e, M. Lalonde, and M. Beaulieu, "Toward an application of content-based video indexing to computer-assisted descriptive video," in *Proc. Canadian Conf. Comput. Robot Vision*, Quebec City, QC, Canada, Jun. 2006, pp. 731–734.
- [17] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. SIGIR*, Berkeley, CA, 1999, pp. 50–57.

- [18] M. Jia, X. Fan, X. Xie, M. Li, and W-Y. Ma, "Photo-to-Search: Using camera phones to inquire of the surrounding world," in *Proc. 7th Int. Conf. Mobile Data Manage.*, Nara, Japan, 2006, pp. 46.
- [19] Y. Ke and R. Suthankar, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM Multimedia Conf.*, New York, 2004, pp. 869–876.
- [20] Y. Ke, R. Suthankar, and L. Hutson, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. CVPR*, Washington, D.C., 2004, pp. II-506–II-513.
- [21] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," in *Proc. IEEE Conf. Multimedia Comput. Syst.*, Florence, Italy, Jun. 1999, pp. 685–690.
- [22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [23] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. BMVC*, Cardiff, U.K., Sep. 2002, pp. 384–393.
- [24] K. Mikolajczyk, and C. Schmid, "A performance evaluation of local descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
- [25] S. Porter, M. Mirmehdi, and B. Thomas, "Video indexing using motion estimation," in *Proc. 14th Brit. Mach. Vision Conf.*, Norwich, U.K., Sep. 2003, pp. 659–668.
- [26] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Proc. ICCV*, Beijing, China, 2005, pp. 883–890.
- [27] P. J. Rousseau, "Least median of squares regression," *J. Amer. Statist. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.
- [28] S. Satoh, M. Takimoto, and J. Adachi, "Video retrieval and annotation: Scene duplicate detection from videos based on trajectories of feature points," in *Proc. Int. Workshop Multimedia Retrieval MIR*, Augsburg, Germany, 2007, pp. 237–244.
- [29] F. Schaffalitzky, and A. Zisserman, "Automated location matching in movies," *Comp. Vision Image Understand.*, vol. 42, no. 2–3, pp. 236–264, Nov.–Dec. 2003.
- [30] F. Schaffalitzky, and A. Zisserman, "Multi-view matching for unordered image sets, or How do I organize my holiday snaps?" in *Proc. 7th Eur. Conf. Comput. Vision*, Copenhagen, Denmark, 2002, pp. 414–431.
- [31] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects categories in image collection," in *Proc. IEEE Inter. Conf. Comput. Vision*, Beijing, China, Oct. 2005.
- [32] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *Int. J. Comput. Vision* vol. 67, no. 2, pp. 189–210, Apr. 2006.
- [33] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. CVPR*, Washington, D.C., 2004, pp. I-488–I-495.
- [34] J. Sivic, and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, Nice, France, 2003, pp. 1470–1477.
- [35] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3-D," in *Proc. ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul.–Aug. 2006.
- [36] W. Tavanapong and J. Zhou, "Shot clustering techniques for story browsing," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 517–527, Aug. 2004.
- [37] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.
- [38] R. Unnikrishnan, and M. Hebert, "Extracting scale and illuminant invariant regions through color," in *Proc. 17th Brit. Mach. Vision Conf.*, Edinburgh, Scotland, Sep. 2006.
- [39] J. Van de Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. ECCV2006*, Part II, Graz, Austria, pp. 334–348.
- [40] J. Vendrig, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 492–499, Dec. 2002.
- [41] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Proc. CVPR*, vol. 2, New York, NY, Jun. 2006, pp. 1597–1604.
- [42] J. Yao and W. K. Cham, "Robust multi-view feature matching from multiple unordered views," *Pattern Recognit.*, vol. 40, no. 11, pp. 3081–3099, 2007.



Maguelonne Héritier received the degree in telecommunications engineering from École CPE, Lyon, France, in 2003 and the M.Eng degree from École de Technologie Supérieure, Montréal, QC, Canada, in 2005.

She joined the Computer Research Institute of Montréal, QC, Canada in 2005. She has worked on different projects related to 3-D volume rendering, machine learning, video and image indexing.



Langis Gagnon (M'00) received the Ph.D. degree in mathematical physics from the Université de Montréal, Montréal, QC, Canada in 1988.

Until 1995, he was a Research Officer with the Center d'Optique, Photonique et Laser of the Université Laval, Laval, QC, the Center de Recherches Mathématiques, the Université de Montréal, and the Laboratoire de Physique Nucléaire, the Université de Montréal. From 1995 to 1998, he was Specialized Researcher at Lockheed Martin Canada, where he worked on developing radar image processing tools

for aerial surveillance applications. He joined the Research and Development Department of the Computer Research Institute of Montréal in 1998, where he is now Principal Researcher and Team Leader (Vision and Imaging). He has published nearly 150 scientific articles related to the fields of image processing, object recognition, and nonlinear optical modeling.



Samuel Foucher (M'04) received the degree in telecommunications engineering from Université de Sherbrooke, Sherbrooke, QC, Canada, and the Ph.D. degree in radar imaging from Université de Rennes I, Rennes, France, in 2001.

Between 1999 and 2002, he was a Research Scientist, working on image mining for the Insat-2E satellite for the India Meteorology Department. He then joined the Computer Research Institute of Montréal, Montréal, QC, Canada, to work on polarimetric image processing and content-based video

technologies. His research interests are in image processing, multi-resolution encoding techniques (wavelets), data fusion, belief theory, and Markovian techniques.