Key-Places Detection and Clustering in Movies Using Latent Aspects

Maguelonne Héritier, Samuel Foucher, Langis Gagnon

R&D Department, CRIM, 550 Sherbrooke West, Suite 100, Montreal, QC, CANADA, H3A 1B9
{maguelonne.heritier, samuel.foucher, langis.gagnon}@crim.ca

ABSTRACT

We describe a new method to find and cluster recurrent key-places in a movie. It consists of an unsupervised classification of shots that are taking place in the same physical location (key-place). Our approach is based on finding links between key-frames belonging to a same key-place. We use a probabilistic latent space model over the possible match points between the image sets. This allows extracting significant groups of local descriptor matches that may represent characteristic elements of a key-place. A preliminary test on a full-length movie gives a recognition rate of 78.0% on the key-places clustering.

Index Terms— Scene categorization, scene matching, content-based indexing, descriptive video, video processing.

1. INTRODUCTION

The aim of this paper is to report about a new approach to detect and cluster recurrent locations in full-length movies using latent space modeling. In most movies, the storyline is taking place in a set of recurrent locations called key-places. Key-places carry important high level semantic information that can be useful to video indexing/retrieval applications. For instance, key-places identification is a very important element for the production of descriptive video. Descriptive video, also known as audiovision, is a narration added to the movie audio track to orally describe visual elements for the blind and seeing-impaired people. This industry is growing due to the imposition of regulations to increase broadcasting of programs with descriptive narration. The work we present here is part of a larger project targeting the development of software tools for computer-assisted descriptive video [4].

Automatic location recognition in movies is a complex problem because of the various scene appearances due to camera viewpoints, foreshortening, scale change, partial occlusion, lighting changes, etc. One approach that has been proposed is based on the use of affine covariant regions ([7], [12]). It uses multiple instances of an object in a shot in order to enable object-based location identification.

Another recent approach that is emerging uses the concept of probabilistic Latent Semantic Analysis (pLSA) ([2][3][11]) which generally addresses unsupervised visual learning problems. PLSA generative models are used in natural language processing and statistical text analysis to discover topics in documents [5]. It is based on the concept of bag-of-words (bag-of-visterms in the visual field) to describe the document (image). This approach has two drawbacks: *polysemy* (i.e. a single visterm that may represent different scene content) and *synonymy* (i.e. several visterms that may characterize the same image content). Probabilistic latent space models have been proposed to capture co-occurrence information between elements in a collection of discrete data in order to raise the ambiguity of the bag-of-words representation. In [2][3][11], image of a scene is represented by a local visterms distribution, denoted as topic (e.g. grass, roads, buildings) obtained by unsupervised learning. This is used to perform scene classification.

There is not much literature specifically regarding classification of film shots in terms of similar physical location. An interesting related work is the one of Schaffalitzky and Zisserman [12] that addresses the problem of finding matches in a collection of images with respect to a query image. They use local invariant descriptors from the wide baseline approach [7] which is a very time consuming process. The invariant features alone are not discriminant enough, which result in many mismatches. Their matching method proceeds in three steps, each using increasingly stronger constraints: (1) matching, using "neighbourhood consensus", (2) local verification of putative matches using intensity registration and cross-correlation and (3) semi-local and global verification where additional matches are grown using a spatially guided search. Those that are consistent with views of a rigid scene are selected using fitting epipolar geometry.

In this paper, we address classification of film shots in terms of similar physical location (key-place) based on a Latent Dirichlet Allocation (LDA) approach [1]. The LDA is a new generative model derived from pLSA. It has been shown to be superior to pLSA because it can be applied to documents that contain several topics and can generate documents not in the training corpus. We use LDA to extract significant matches distribution over the possible matches between key images on a film. This generative model provides a discrete discriminant analysis over matches. The visterms are seen as a group of local descriptors that match

together. Visterms distribution is seen as part of "topic" which is in fact a typical element representation from a scene in a higher semantic level (Figure 1).

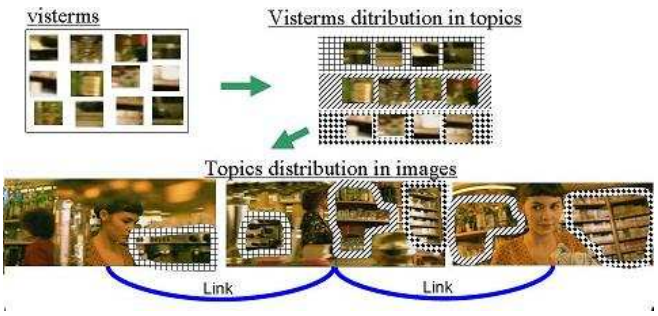

Figure 1. Principle of latent semantic analysis

The paper is organized as follows. Section 2 defines the general concept of our method. Section 3 describes our experiment settings and Section 4 presents our results.

2. METHODOLOGY

2.1. Image set representation

The construction of the bag-of-visterms (BOV) is done from a set of several key-frames extracted for each movie shot. First, regions of interest (ROI) are automatically detected in the image with the difference of Gaussians (DOG) point detector over which one computes local descriptors using the Scale Invariant Feature Transform (SIFT) [6]. We use SIFT because it performs the best in terms of specificity of region representation and robustness to image transformations [8].

Second, in order to obtain a text-like representation, descriptors must be quantized. Unlike previous approaches, we do not use a K-mean clustering for descriptor quantization ([2][3][11]). Indeed, we do not want to generalize a descriptor in order to avoid mismatch. Instead, we use K-nearest neighbour (K-NN) between SIFT descriptors belonging to different images to create the visterms. A visterm is a set of local descriptors that match together. We call *vocabulary* the set *V* of all visterms. Finally, the BOV representation *h* is constructed from the local descriptors according to

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

where $n(d, v_i)$ is the number of occurrences of visterm v_i in a sub-image *d* and N_V is the size of the visterm set. This representation contains no information about the spatial relationships between visterms, the same way that the standard bag-of-words text representation removes the word ordering information. This is why we use sub-image instead of the entire image. Each image is divided into several sub-images of different size (with or without overlapping).

2.2. Generative model

Following the pLSA framework, we have *sub-images* as *documents* and we want to discover *topics* as *semantic characteristics of location* (e.g. cigarette shelves from a bar, tapestry from a room, etc.) so that an image containing instances of semantic characteristics of location is modelled as a mixture of topics. The models are extracted from sub-images by using the *visterm* analogue of a *word*, formed by SIFT matching feature descriptors. Suppose a collection (corpus) of sub-images $D = d_1,..,d_{N_D}$ with visterms from a visual vocabulary $V = v_1,..,v_{N_V}$. One can summarize the data in a $N_D \times N_V$ co-occurrence table (BOV) of counts $h_i(d_j) = n(d_j, v_i)$. In pLSA, there is also a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = z_1,..,z_{N_Z}$ with each observation. A joint probability model *P(v,d)* over $N_D \times N_V$ is defined by the following mixture

$$P(v \mid d) = \sum_{z \in Z} P(z \mid d) P(v \mid z) \qquad (2)$$

where $P(v \mid z)$ are the topic specific distributions and each image is modelled as a mixture of topics, $P(z \mid d)$. (See [5] for a detailed explanation of the model) .

The LDA is a corpus generative model [1]. Documents are represented as random mixtures over latent topics. The framework treats the topic mixture weights as a k-parameter hidden random variable (θ) and places a Dirichlet prior on the multinomial mixing weights. The model parameters are estimated using the maximum likelihood principle, using a set of training sub-images *D*. The optimization is conducted using an alternative variational Expectation-Maximization (EM) algorithm. By using an approximation inference algorithm, these independent sub-images 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)*.

3. IMPLEMENTATION

After an automatic shot transition detection step, we summarize each shot in the film using few representative frames (key-frames). To this aim, we use a simple method proposed in [10] based on camera motion estimation to compute overlap between images. The final selection of the best set of images is seen as a shortest path problem (see [10] for details). Small shots that are less than ten frames in

size are eliminated. SIFT descriptors are then calculated for feature points found in the key-frames.

The next step consists in quantizing all those descriptors with a K-NN. A first NN algorithm is used to match descriptors between two different images. Bad matches are removed when their histogram intersection distance are above 0.25 and when the distance to the first nearest neighbour d_{fnn} is above 90% of the distance to the second nearest neighbour d_{snn}

In order to avoid all descriptors to be matched together and form a unique visterm, we divide our initial image set into several random smaller sets of images (height images size in our tests). Then we apply LDA on each BOV created with sub-images extracted from each subset of images. In our tests, we use two sets of sub-images per image made from a grid decomposition. We use $w_{subimage1} = w_{image}/3$ and $h_{subimage1} = h_{image}/2$ for the first set, and $w_{subimage1} = w_{image}/4$ and $h_{subimage1} = h_{image}/3$ for the second. These sub-images 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.

The number of topics is a parameter we need to choose for the LDA algorithm. The topic initialization is done by assigning a random group of visterms from a document to a topic until the maximum number of topics is reached, or when no more visterms are available. We set the maximum number of topics to six.

After the LDA application, we select the best document and visterms for each topic. When two of the selected documents share more than three visterms, a topic link is formed. A further step is added to filter out wrong links. It consists in eliminating topic links for which visterm SIFT descriptor matches are not within the same range of scale and direction variation.

4. TEST

We have tested our method on the French feature-length movie "Le fabuleux destin d'Amélie Poulain" from which we automatically extracted 1,223 shots and 1,561 key-frames. The four main places of the film are the bar (20% of the shots), the neighbour's apartment (11%), Amélie's apartment (8%) and the grocery (4%).

4.1. Shot links

We have observed that very subtle good matches can be generated between images using the first NN algorithm on SIFT descriptors. The LDA approach is able to separate those subtle matches from the large number of matches generated. 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. Figure 2 shows examples of links for three key-places in the film.

Figure 2. Links for "Amélie's kitchen" class (top), "bar" class (middle) and "grocery" class (bottom). Matches are in white boxes.

False links usually appear in common similar visual structure, like striated or squared structure. They are often composed of less than five visterms. Figure 3 shows examples of false links.

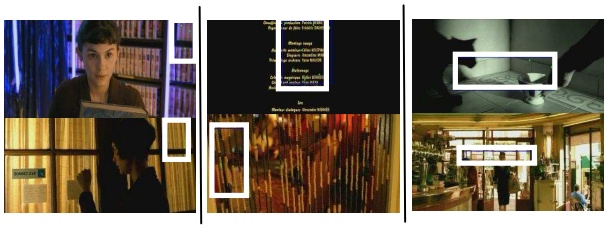

Figure 3. Examples of wrong links.

4.2. Shot clustering and key-places extraction

Grouping of key-frames belonging to a same key-place is performed by constructing a graph between shots using links extracted by the LDA. Clusters (sub-graphs) have been identified using a spectral clustering algorithm [9]. 822 out of the 1,223 shots are considered as a part of one of the 35 key-places defined in the Ground Truth (GT). We evaluate the link performance by calculating, before spectral clustering, the rate of wrong links between shots which gives 8 %. The spectral clustering is set to extract homogeneous clusters but may produce too much clusters. 32 of the 35 key-places in the GT are represented in one or several

clusters. For each key-place, we evaluate the Recognition Rate (RR) and the False Alarm (FA) rate. Table 1 shows measures for the four top places in the movie. A place can be represented by several key-places (e.g. kitchen and bedroom for Amélie's apartment). Amélie's apartment is represented by five clusters in the GT.

Places	# Shots	RR	# Clusters GT	# Clusters Obs.	Top Cluster RR	FA
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 1. Performance measures. "# Shots" is the number of shots taken in a specific place. "# Clusters GT" is the number of key-places identified in the GT for a specific place. "Clusters Obs" is the number of clusters observed for a specific place. "Top Cluster RR %" is the recognition rate of shots within the dominant observed cluster for a given place.

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 apartment", "Neighbour's apartment" and "bar" classes have a large variability because of their many point of views. Also, the place sometime appears in several close-up plans. Only few interactions between actors happen in Amélie's apartment. Also, this place often appears in several short shots and in close-ups. Despite of that, the global recognition rate is quite satisfying (78%) with a false alarm rate of 0.01% which is important for the user-labelling of the clusters.

5. DISCUSSION AND CONCLUSION

The more often a particular location appears in the movie, the more this key-place is susceptible to present meaningful details that will enable LDA to detect its content as a predominant topic. This process is similar to what humans do when they assimilate key-places in a movie. If a location is shown several times and has a lot of specific details, the viewer quickly considers it as a reference location.

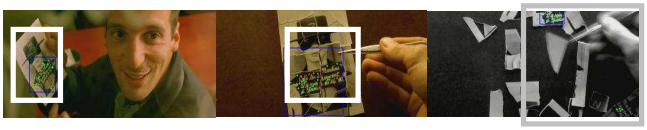

Figure 4. Key-object examples: photography.

We notice that the process of extracting specific visual structure works as well as for key-faces and key-objects (Figure 4). However, this can be problematic for the key-place clustering task when the face of the principal actor is present in different locations. In fact, LDA can then assign the faces as a topic link between theses locations. We resolve this problem in part by separating location topics from face topics using information derived from a face detector previously applied to the movie.

In conclusion, we have presented the first results of a new method to automatically cluster recurrent key-places in a movie. It is based on a probabilistic latent space model over the possible local descriptor matches between the set of key-frames of the movie shots. The method is able to extract groups of significant matches that may represent a semantic characteristic relative to a key-place.

At this early point, the approach gives promising results. Future work will consist in parameter optimization through additional tests on a large bank of videos and on the addition of pre-processing steps. In another application area, we could exploit the technique to improve face clustering in a movie, since the LDA algorithm allows links between face views which are difficult to extract with classical face detection (e.g. in very close-up shots).

ACKNOWLEDGEMENTS

This work is supported in part by (1) the Department of Canadian Heritage (www.pch.gc.ca) through the Canadian Culture Online program and (2) the Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE) of the Gouvernement du Québec.

REFERENCES

[1] D. M. Blei, A. Y. Ng, M. I. Jordan. "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003

[2] A. Bosch, A. Zisserman, X. Munoz, "Scene Classification via pLSA", ECCV, 2006

[3] L. Fei-Fei, P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", CVPR, 2005

[4] L. Gagnon, S. Foucher, F. Laliberté, M. Lalonde, M. Beaulieu, "Toward an Application of Content-Based Video Indexing to Computer-Assisted Descriptive Video", CRV, 2006

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing", SIGIR, 1999

[6] D. G. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints", IJCV, 2004

[7] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust Wide Baseline Stereo From Maximally Stable Extremal Regions", BMVC, pp. 384-393, 2002

[8] K. Mikolajczyk, C. Schmid "A Performance Evaluation of Local Descriptor", PAMI, Vol. 27, pp. 1615-1630, 2005

[9] A. Y. Ng, M. Jordan, Y. Weiss. "On Spectral Clustering: Analysis and an Algorithm", NIPS, 2002

[10] S. Porter, M. Mirmehdi, B. Thomas. "Video Indexing using Motion Estimation", ECCV, 2006

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, L. Van Gool. "Modeling Scenes with Local Descriptors and Latent Aspects", ICCV, 2005

[12] F. Schaffalitzky, A. Zisserman, "Automated Location Matching in Movies", CVIU, Vol. 42, pp. 236, 264, 2003