CRIM Notebook Paper - TRECVID 2008
Video Copy Detection Using Latent Aspect Modeling Over SIFT Matches

M. Héritier, S. Foucher, *Member, IEEE,* L. Gagnon, *Member, IEEE*

R&D Department, Computer Research Institute of Montreal (CRIM)
550 Sherbrooke West, Suite 100, Montreal (Quebec), CANADA, H3A 1B9

Abstract

Approach we have tested in our submitted runs. Our approach consists in finding links between video shot key-frames, based on the use of a probabilistic latent space model over local matches between the key-frame images. This allows the extraction of significant groups of local matching descriptors that may represent common characteristic elements of near duplicate key-frames. This is combined with various pre-processing steps designed to accelerate and improve the matching process for any query type, as well as post-processing steps designed to accurately find the copied video segment borders. We have submitted 3 runs. The first run (*Run1*) uses an automatic cropping process, a local descriptor filter and a global characteristic filter as part of the pre-processing phase. A RANSAC-based post-processing step is applied on the time code of the detected key-frames copy. A video insert detector was added for the second run (*Run2*). The third run (*Run2Faster*) is the same as the second run with the use of smaller images.

Differences we found among the runs: No significant differences were found amongst the three different runs. Unfortunately, a file manipulation error resulted in the second and the third run being deprived of receiving the correct video insert filter detection output. Therefore, these runs received no video insert detection response as input, causing some video insert copy detections undetectable.

Relative contribution of each component of our approach: The probabilistic latent space model over local matches between the key-frame images produces a fast, robust and accurate filtering process in relation to all possible local matches. This approach works well even if there are only a few local matches between the key-frames of the copied video in question. Therefore, only a limited number of local descriptors are necessary, resulting in a more robust copy detection process. Unfortunately, the large number of local matches still makes the process rather time consuming.

What we learned about runs/approaches and the research question(s) that motivated them: Approaches based on local descriptor matching are efficient for the copy detection task. It is robust to many transformations. However, these approaches are not efficient for some query categories. For instance, the flipped query type provides totally different local descriptors and the insert query requires the use of local descriptors at a higher-resolution level. Also, local descriptor matching is very time consuming. It would then be interesting to combine local descriptor matching approaches with global characteristic comparison approaches to provide a prior knowledge about the query type. Global characteristic comparison approaches alone are not efficient enough for accurate copy detection.

I - Introduction

Near-duplicate detection (NDT) in movies is a relatively new topic ([29],[20],[21],[16],[11]) as it offers an alternative to watermarking for copyright control, business intelligence, advertisement tracking and law enforcement investigations. NDT often proceeds via a video summarization approach like reducing a video in a set of key-frames. The copy detection task then consists in finding near-duplicates in key-frame images

([17],[39],[40]). Matching key-frames through a set of key-points is an interesting strategy because it is robust to occlusions and illumination changes. Moreover, invariant descriptors for the key-points provide robustness 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],[4],[17],[31],[35],[38]) consists in filtering out the outliers between the whole key-frames using robust matching methods such as RANSAC or Least Median of Squares (LMS). However, those fitting methods perform poorly when the ratio of inliers falls below 50%. This means that a large overlap is required between a pair of images for an efficient matching process. In practice however, key-frames of two similar video segments can differ significantly due to the presence of motion in the scene or the key-frame generation process. Also, RANSAC is not efficient if there are only few inliers between 2 near duplicate key-frames. In [23], Lowe proposed to cluster features within 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 seek to find common spatial patterns (e.g. [30],[31],[32],[33]). These 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, it is always possible to obtain erroneous matches due to the presence of common local structures. Some authors ([32],[33]) use an efficient representation inspired from text analysis called "Bag Of Words" (BOW), in order to represent neighborhoods. BOW consists in representing a text document as a vector; counting the number of occurrences of different words as features. In [32],[33], descriptors are quantized into clusters which are analogous to "words" in a text document. The BOW representation has two shortcomings when dealing with ambiguities: *polysemy* (i.e. a word that has two different meanings) and *synonymy* (i.e. two words with 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. The probabilistic Latent Semantic Analysis (pLSA) is one of the best known BOW models [14]. BOW generative models are used in natural language processing and statistical text analysis to discover topics in documents [14]. They have recently been applied to classification in the image processing field ([7],[8],[27]). Local patches are called visterms, and are modeled as basic building blocks of an image; analogous to words in text documents. Images are represented as a collection of visterms.

In ([12],[13]), BOW generative models are used to extract and link place features and cluster recurrent physical locations (key-places) within a movie. It 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. This allows the extraction of significant groups of matching descriptors that may represent characteristic elements of a key-place. Here, we adapt this approach for the video copy detection task. The BOW is used to represent key-frame images. BOW generative model filters out uninformative matches, generated by very common image structures, and extract groups of matches that may represent structural elements representative of near duplicate key-frames. Inliers are extracted, whatever the outlier number, by using a latent value for each match. A latent value is a context value shared by a group of local matches that may represent a structural characteristic element (analogous to a "topic" for text document). We use the Latent Dirichlet Allocation (LDA) generative model [2], which is a new model derived from pLSA [14], to extract significant matches' distribution between key-frames. This generative model provides a discrete discriminant analysis over matches. The visterms are seen as a group of local descriptors that match together. The significant extracted visterm distributions are seen as part of latent "topics" which are, in fact, typical structural elements of a key-frame. 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.

This notebook paper is organized as follows. Sections II, III and IV give details about the methodology and implementation of our approach for the video copy detection task. Section V presents the evaluation process and performance results on the TRECVID dataset 2008.

II - Pre-processing

Automatic crop

This step removes the eventual black borders from the movie images and normalizes the video size. The automatic crop removes the horizontal and vertical border when the maximum pixel value and the maximum difference between pixels values fall below respective thresholds. This simple approach works most of the time but there are still some videos were the borders remain undetected.

Key-frame extraction

Once the automatic shot transition detection is completed, each shot is then summarized in a few representative frames (key-frames). To this aim, we compute the overlap between images using a simple method based on camera motion estimation [26]. The algorithm finds the optimal frame path over the shot which then minimizes the overlap between frames. When there is a potential insert, we calculate the optimal frame path using the maximum absolute value of the camera motion estimated on the potential insert and on the entire video.

Local descriptor extraction

We extract local descriptors for each key-frame. First, Regions Of Interest (ROI) are automatically detected in the image with a difference of Gaussians (DOG) point detector from which we derive local descriptors using SIFT [23]. We use SIFT because it performs the best in terms of region representation specificity and robustness to image transformations [25]. We have tried several other descriptors (MSER based [24], PCA-SIFT [18], etc.) in a preliminary work [9] but found that SIFT with DOG provided the best combination to establish the difficult correspondences between images of various appearance and quality (low illumination, night, smoothing filter, etc.).

Descriptor filtering

In order to accelerate the linking process, we need to deal with the fewest possible number of local descriptors. We could simply reduce the key-frame image resolution but this would alter the accuracy of the image representation. Another idea consists in eliminating the more common local descriptors which are not discriminative enough. For instance, local descriptors corresponding to straight lines or corners can be found in many images. This type of local descriptor is not specific enough to accurately describe an image and often generates erroneous matches. We build a database of “good” and “bad” local descriptors from a set of images containing redundancy. We generate local matches between images using K-Nearest Neighbors (K-NN). The “good” descriptors are those with their first nearest neighbor distance below 0.7 times the second nearest neighbor distance. The “bad” descriptors are those with their first nearest neighbor distance above 0.9 times the second nearest neighbor distance. Indeed, it is harder to find a similar discriminative descriptor in the same image while it is easier to find a similar descriptor when it is a very common one. We use an Adaboost classifier on some features extracted from the SIFT signatures which are the entropy and the sum of SIFT bins corresponding to different direction ranges (0 degrees, 90 degrees, out of the 0 and 90 degrees bins). Classification results on our database give a recall rate of 85% for the “good” descriptors and a recall rate of “63%” for the “bad” descriptors.

Insert detection

Inserts are detected by accumulating local gradient intensities over time. The main hypothesis is that insert edges are stationary whereas gradient values within the scene vary over a sufficient number of frames. However, this approach will not perform well for videos with low motion activity with a fixed background. Also, false alarms can occur in presence of overlaid texts and logos.

Global features comparison

Global features comparison is not sensitive enough to accurately detect video copies because they only offer a coarse representation of the video. However, it can help to discard candidates when the global features are different enough, prior the application of a more sensitive comparison measure with local descriptors. Global feature comparison is fast; therefore it is an interesting way to accelerate the copy detection process. In our experiment, we have only used the edge orientation histogram as a global feature.

The edge orientation histogram is quite robust to many transformations like illumination changes, small view point changes, small insertion, crop, shift, noise, etc. During the preprocessing step, we calculate the edge orientation histogram for each key-frame of the reference video, as well as for their flipped copy. We also calculate the histogram for each key-frame of the query video and one for each potential video insert in each key-frame. When the global feature of the flipped copy is closer than the global feature of the non-flipped version, we use the local descriptors extracted from the flipped copy. When the global feature from the insert is closer than the global feature from the entire key-frame, we use the local descriptors extracted from the insert at a higher resolution level.

III - Key-frame link extraction

We extract groups of local matches between near duplicate key-frames. We use the concept of "Bag of visterms" (BOV) for representing each key-frame, in conjunction with a different method of building the representations based on K-NN. We then apply a generative probabilistic model to extract groups of local matches that represent a common structure representative of 2 near duplicate key-frames.

Image set representation

The construction of BOV is done from a set of several key-frames. However, we had to choose a very weak selection step on SIFT features' matches in order to conserve difficult but important correspondences. It became apparent that high-quality correspondences are hidden among a large amount of outliers. We configured the algorithm to privilege a high correspondence rate over outliers so that it results in a maximum number of correspondences within key-frames from the query video and key-frames from the reference video. Second, in order to obtain a text-like representation, descriptors must be clustered. We do not use prototype-based clusterings such as K-means for descriptor quantization ([8],[27],[32],[33]). 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 prototype may be not well defined. Also, the cluster prototypes 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 within images from query video and images from reference video based on the Manhattan distance. However, conventional K-NN is computationally expensive. We rather use an approximate K-NN approach based on a priority tree search [1]. Bad matches are discarded when their distances are above 0.6, and when the distance to the first nearest neighbor is above 90% of the distance to the second neighbor. Finally, a BOV representation h is built 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 visterm occurrences v_i in an 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.

Generative model

We describe here the probabilistic latent space modeling applied over the possible local matches within key-frames image from the query videos and within key-frames image from the reference videos. We propose it as an efficient way to filter out the outliers. We only give a brief overview of the main concepts of the pLSA (probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) models applied to visual content.

According to the pLSA framework [14], we have images represented as documents and we want to discover topics as common structural characteristics of near duplicate key-frames such that 2 images sharing instances of common structural characteristics are modeled with shared topics. Those shared topics form a topic link. The models are extracted from the BOV representation of images. The visterm analogue to a word is formed by SIFT matching feature descriptors. Let a collection (corpus) of sub-images $D = d_1,..,d_{N_D}$ with visterms from a visual vocabulary $V = w_1,..,w_{N_V}$. 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 [14] 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

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

where $P(w \mid z)$ are the specific topic distributions. Each image is modeled as a mixture of topics $P(z \mid d)$ [14].

In pLSA, each document, represented as a list of numbers (the mixing proportions for topics), was not 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 [2]. LDA allows each document to exhibit multiple topics with different proportions, and thus, 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*).

Implementation considerations

The larger the set of images is, the larger the number of possible connections will be between the SIFT descriptors and thus among 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 to be quantized to a unique visterm, we have to limit the size of the image set. We then randomly divide the initial query image set into several smaller image sets (7 images in each of our tests). For each reference video, we apply several LDA filtering on those smaller images sets and on all the key-frame images from each reference video.

LDA requires setting up the number of topics. This is set automatically during the initialization step. A random topic is attributed to each visterm which are spread over many documents. 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. Each key-frame image is described by its BOV. After the LDA application, we select the best images and visterms for each topic. Then, a topic link is formed when two of the selected images share more than four selected visterms. A further step is added to filter wrong links. It consists in eliminating topic links for which the SIFT matches are not within the same range of scale variation.

IV - Post-Processing

Copied video segment extraction

Copied video segments are detected once links are extracted between the query and the reference key-frames. We apply RANSAC in the temporal domain in order to estimate the time shift and dilation between the times codes of the detected links. This step ensures that detected links are forming a coherent segment in time up to a translation and scaling factors. Finally, the shot boundaries from which the selected near-duplicate key-frame belongs to, define the time range of the near duplicate video segment.

Confidence value

The confidence value is calculated from the number of local matches first extracted by the probabilistic latent space model and then selected by the video copy segment RANSAC estimation. For each query, this number is divided by the maximum number of local matches found for this query and is multiplied by 100. In order to not penalize the copy candidates with many local matches, we add 1/10 of the initial number of local matches.

V – Results

Detection curves

Table 1, Table 2 and Table 3 summarize the DET curves results in terms of miss and false alarm rates provided by the TRECVID evaluation tools [34].

Run id	1	2	2Faster
Miss rate	0.471	0.4672	0.468
FA (Events/Hour)	0.4617	0.506	0.4777

Table 1: Mean Miss rate and mean False Alarm rate for each Runs (all transforms included)

Transform nb.	1	2	3	4	5	6	7	8	9	10
Miss rate	0.31	0.79	0.32	0.32	0.18	0.36	0.44	0.50	0.73	0.72
FA (Events/Hour)	0.49	0.49	0.55	0.76	0.37	0.58	0.45	0.43	0.29	0.40

Table 2: Mean Miss rate and mean False Alarm (FA) rate for each transform (all Runs included)

Transform nb.	1	2	3	4	5	6	7	8	9	10
Miss rate	0.27	0.69	0.28	0.28	0.16	0.31	0.34	0.39	0.59	0.67
FA (Events/Hour)	0.38	0.32	0.34	0.54	0.28	0.30	0.32	0.32	0.14	0.20

Table 3: Best Miss rate and best False Alarm (FA) rate for each transform (among all runs)

Video copied segment are relatively well detected for transforms 1, 3, 4 and 5; the mean miss rate for all runs is between 18% and 32%. The mean false alarm rate is above 0.79 events/hours for all transforms. The miss rate for transform 2 is_high (79%) because a file manipulation error yielded that the second and third run did not get the right video insert filter detection output. Therefore, these runs got a no video insert detection response as input so video insert copy detections were almost not detected. Quality degradation transforms (transformations 6 and 7) give a mean miss rate between 36% and 44% while post-production transforms (transformations 8 and 9) give a mean miss rate between 50% and 73%.

VI - Conclusion

The probabilistic latent space model over local matches between key-frames allows a fast, robust and accurate filtering process among all possible local matches. It works well even if there are only few local matches between the key-frames of the copied video. Therefore, only a limited number of local descriptors are necessary. However, the large number of local matches still makes the process rather time consuming.

We think the proposed approach has a good potential for video copy detection because it is relatively robust to many transformations. However, this approach needs improvements for some query categories. For instance, the flipped query type provides totally different local descriptors and the insert query requires the use of local descriptors at a higher resolution level. Also, local descriptor matching is very time consuming. It would then be interesting to combine local descriptor matching approaches with global characteristic comparison approaches to give a prior knowledge about the query type. Also, it will be interesting to investigate other video representations than static key-frame based representations (e.g. Space Time Interest Points [19]).

References

[1] 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. of the ACM 45*, 6, pp. 891-923, 1998.

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

[3] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features", *International Journal of Computer Vision*, August 2007 (to appear).

[4] M. Brown and D. G. Lowe, "Unsupervised 3D object recognition and reconstruction in unordered datasets", *5th International Conference on 3D Imaging and Modeling*, pp. 56-63, 2005.

[5] M. Brown, R. Szeliski and S. Winder, "Multi-image matching using multi-scale oriented patches", *International Conference on Computer Vision and Pattern Recognition (CVPR2005)*, pp. 510-517.

[6] G. Csurka, C.R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints", in *Proc. European Conference on Computer Vision*, 2004.

[7] 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", *Workshop on Generative-Model Based Vision*, 2004.

[8] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories", *CVPR*, 2005..

[9] S. Foucher, M. Héritier, 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", Technical Report , CRIM-06-05/11, May 2006

[10] L. Gagnon, S. Foucher, F. Laliberté, M. Lalonde and M. Beaulieu, "Toward an application of content-based video indexing to computer-assisted descriptive video", *Canadian conference on Computer & Robot Vision (CRV)*, Quebec City, June 2006.

[11] A. Hampapur and R. Bolle. "Comparison of sequence matching techniques for video copy detection". *In Conference on Storage and Retrieval for Media Databases, pages 194--201, 2002.*

[12] M. Héritier, S. Foucher, L. Gagnon, "Key-Places Detection and Clustering in Movie Using Latent Aspects", *ICIP, San Antonio, 2007.*

[13] M. Héritier, L. Gagnon, S. Foucher, "Places Clustering of Full-Length Film Key-frames using Latent Aspects Modeling over SIFT Matches", *IEEE Trans. on Circuits and Systems for Video Technology (to appear).*

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

[15] M. Jia, X. Fan, X. Xie, M. Li and W-Y. Ma, « Photo-to-Search: using camera phones to inquire of the surrounding world », *7th International Conference on Mobile Data Management*, 2006.

[16] K. Kashino, T. Kurozumi and H. Murase, ``A Quick Search Method for Audio and Video Signals Based on Histogram Pruning", *IEEE Transactions on Multimedia, Vol. 5, No. 3, pp.348-357, Sep. 2003.*

[17] Y. Ke and R. Suthankar, "Efficient near-duplicate detection and sub-image retrieval", *in ACM Multimedia Conference*, 2004, pp. 869-876.

[18] Y. Ke, R. Suthankar and L. Hutson, "PCA-SIFT: a more distinctive representation for local image descriptors", in Proc. of CVPR, 2004.

[19] I. Laptev, T. Lindeberg. “Space-time interest points”. *In International Conference on Computer Vision, page 432-439, 2003.*

[20] J. Law-To, O. Buisson, V. Gouet-Brunet, N. Boujemaa, “Robust voting algorithm based on labels of behavior for video copy detection*”. ACM Multimedia 2006: 835-844*

[21] J. Law-To, L.Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, F. Stentiford*,* “Video Copy Detection: a Comparative Study*, Proceedings of the 6th ACM international conference on Image and video retrieval, p 573-580, July 09-11, 2007, Amsterdam, The Netherlands*

[22] R. Lienhart, S. Pfeiffer and W. Effelsberg, “Scene determination based on video and audio features”. *Proc. IEEE Conf. on Multimedia Computing and Systems*, Florence, Italy, pp. 685 - 690, June 1999.

[23] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, *IJCV*, 2004.

[24] J. Matas, O. Chum, M. Urban and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions”, *BMVC*, pp. 384-393, 2002.

[25] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptor”, *IEEE Trans. on PAMI*, Vol. 27, pp. 1615-1630, 2005.

[26] S. Porter, M. Mirmehdi and B. Thomas, “Video indexing using motion estimation”, *ECCV*, 2006.

[27] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars and L. Van Gool, “Modeling scenes with local descriptors and latent aspects”, *ICCV*, 2005.

[28] P. J. Rousseau. “Least median of squares regression”, *J. Amer. Statist. Assoc.* 79(388), pp.871–880, 1984.

[29] S. Satoh, M. Takimoto and J. Adachi, “Video retrieval and annotation: scene duplicate detection from videos based on trajectories of feature points,” in *Proc. of the International workshop on Workshop on multimedia retrieval MIR 2007*.

[30] F. Schaffalitzky and A. Zisserman, “Automated location matching in movies”, *CVIU*, Vol. 42, pp. 236, 264, 2003.

[31] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets, or “How do I organize my holiday snaps?”, *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[32] J. Sivic and A. Zisserman, “Video data mining using configurations of viewpoint invariant regions”, *CVPR*, 2004.

[33] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos”, *ICCV*, 2003,

[34] A. F. Smeaton, P. Ove and W. Kraaij, “Evaluation campaigns and TRECVID”, *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 321-330. Santa Barbara, California, U&SA*

[35] N. Snavely, S. M. Seitz and R. Szeliski, “Photo tourism: Exploring photo collections in 3D” *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, 25(3), pp. 835-846, 2006

[36] W. Tavanapong and J. Zhou, “Shot clustering techniques for story browsing” *IEEE Trans. On Multimedia*, 2004.

[37] G. Wang, Y. Zhang and L. Fei-Fei. “Using dependent regions for object categorization in a generative framework”**,** *IEEE Comp. Vis. Patt. Recog.* 2006.

[38] J. Yao and W.K Cham. “Robust multi-view feature matching from multiple unordered views”. *Pattern Recognition*, Vol. 40, Issue 11 pp. 3081-3099, 2007.

[39] D.-Q. Zhang and S.-F. Chang, “Detecting Image Near-Duplicate by Stochastic Attribute Relational Graph Matching with Learning”, ACM Multimedia 2004

[40] W. Zhao, C.-W. Ngo, Hung-Khoon Tan and Xiao Wu, “Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning”. *IEEE Transactions on Multimedia 9(5): 1037-1048 (2007)*