SEMANTIC CONCEPTS FOR CONTENT FILTERING ON VIDEO SHARING SOCIAL NETWORKS

Antonio da Luz
Computer Science Department (DCC) - Federal University of Minas Gerais (UFMG)
Belo Horizonte – MG - Brazil

Eduardo Valle
School of Electrical and Computer Engineering (FEEC) – State University of Campinas (UNICAMP)
Campinas – SP – Brazil

Arnaldo de A. Araújo
Computer Science Department (DCC) - Federal University of Minas Gerais (UFMG)
Belo Horizonte – MG - Brazil

ABSTRACT

In this work we are concerned in use semantic concepts to filter spam videos on video sharing social networks. Specifically, we investigate how much semantic-based information analysis, based on content-based visual information retrieval (CBVIR), can aid in detecting spam videos. This is a very challenging task, because of the high-level semantic concepts involved; of the assorted nature of social networks, preventing the use of constrained a priori information. In addition, a spam video is by nature context-dependent, forcing us to take into account the context of the videos within their threads in the classification. Web services for the sharing of clips of video are extremely popular; and the recent boom of intelligent mobile devices such as smartphones and tablets equipped with fast network access and good-quality cameras and displays has reduced the interval between content creation and broadcasting. That proliferation of new content, and the immediacy of its availability, is not without problems, creating a demand for mechanisms to control abuses and terms-of-use violations. We propose an approach based on bags-of-topic-differences, which improves considerably over the use of the baseline bags-of-words model, by allowing us to incorporate the context of the video in the representation. Our model is evaluated in challenging video dataset, showing very encouraging results.

KEYWORDS

Semantic Video Classification, Latent Semantic Analysis, Bag-of-Features, STIP.

1. INTRODUCTION

Online communities built upon the production, sharing and watching of short video clips have been fostered by the popularization of broadband web access and the availability of cheap mobile video devices. The crowds of users who employ the services of websites like Dailymotion, MetaCafe and YouTube, not only post and watch videos, but also share ratings, comments, “favorite lists” and other personal appreciation data. Nowadays it is not uncommon for video to be acquired and edited in a single device, and immediately uploaded to be shared with friends and colleagues.

That proliferation of new content, and the immediacy with which it is broadcast, is not without problems. The emergence of video services has created a demand for specialized tools, including mechanisms to control abuses and terms-of-use violations. Indeed, the success of social networks has been inevitably accompanied by the emergence of users with non-collaborative behavior, which prevents those services from operating evenly. Those behaviors include instigating the anger of other users (“trolling”, in the web jargon), diffusing materials of genre inappropriate for the target community (e.g., advertisement or pornography in inadequate channels), or manipulating illegitimately popularity ratings.

Non-collaborative behavior pollutes the communication channels with unrelated information, and prevents the virtual communities from reaching their original goals of discussion, learning and entertainment. It alienates legitimate users and depreciates the social network value as a whole (Deselaers et al., 2008).

The research on CBVIR was started as research on feature-based visual information retrieval, with impressive, but not complete, success (see Zhang, 2007 for an assessment). The limitations faced by CBVIR can be attributed to the much discuses "semantic gap", which is the lack of coincidence between the low-level information automatically extracted from visual data (e.g. color and texture descriptors) and the high-level interpretation a user would give to that same data (Smeulders et al., 2000). That motivates the efforts of semantic-based visual information retrieval (SBVIR): how to bridge the gap between the perceptual descriptions and the semantic meanings.

Many approaches boarding the semantic gap problem in different content-based retrieval tasks ((Yang et al., 2007), (Benmokhtar and Huet, 2007), (Wei and Ngo, 2008)). Generally, is followed a common way, establishing a set of semantic topics (in a specific domain or not) and use a large dataset to train a classifier to recognize each one.

In this paper, we are concerned with the detection of a kind of non-collaborative behavior, spam videos. To demonstrates the performance of our proposed methodology we use a specific feature of a popular social network, video answers in Youtube. Specially, we are interested on detecting spam in threads of video answers, where the user can post a video in response to another. Here, we consider spam as a video answer whose subject is unrelated with the original video (sometimes advertisement, commercial or not; sometimes videos posted in the hope to attract attention; sometimes videos posted intentionally to anger the other users). Figure 1 illustrates the diversity of the spam phenomenon.

Our work will contributes directly in the reduction of achievement of unwanted videos in results of search of videos of a given topic of interest. Specially regarding to users that use smartphones to access these networks, our methodology enables the development of a lot of new applications. The 3G access is yet very expansive and charged based on volume of exchanged data.

Perhaps the most serious problem to detect spam in video threads is its relative, context-dependent, nature. Accepting the definition that a spam video is simply a video unrelated to the thread topic, the same video (e.g., a viral video with a celebrity breakdown) may be spam in a thread (e.g., a thread about how to cook asparagus correctly), but not in another (e.g., a thread about famous people behaving oddly). It is, of course, possible that some videos are intrinsically more probable to be used as spam (like viral videos) than others (like someone cooking asparagus), but this does not solve, by itself, the problem.

We also face the semantic gap, but the specific problem addressed in this work requires that be adopted a different strategy. Our target is not the classification in predefined semantic topics, but discovers the presence of a non-legitimate element between a set of legitimate ones. This problem can be treated as an odd-man-out problem. This is justified by the fact that in the video answer thread there are many different videos, but all legitimate ones has related content with the addressed theme in the thread. And, the spam videos do not have related information. The related content is defined considering a semantic level, not necessarily the visual content is the same.

The other serious difficulty is the large variety of visual content that can be found in legitimate elements. Even considering a very restricted thread (e.g., how to cook asparagus) and observing only the legitimate answers, the diversity of the videos is overwhelming. Even a human operator has sometimes difficulty in establishing the legitimacy × spam status of the videos by watching just the images.

Finally, we face other difficulties, dictated by the flexible nature of the social networks. The number of videos in the threads is not fixed, and is usually quite small. This makes more difficult to establish a model of visual content that characterizes the legitimate videos in that specific thread. And, this knowledge can not be availed in another answer thread.

Considering these situations, the attempts to identify the spam videos must consider the contextual location of each video. Another consideration is in respect to the visual complexity of the videos in a thread. In addition of the explicit visual information, it is important to consider the occurrences of non-explicit visual contents relationships.

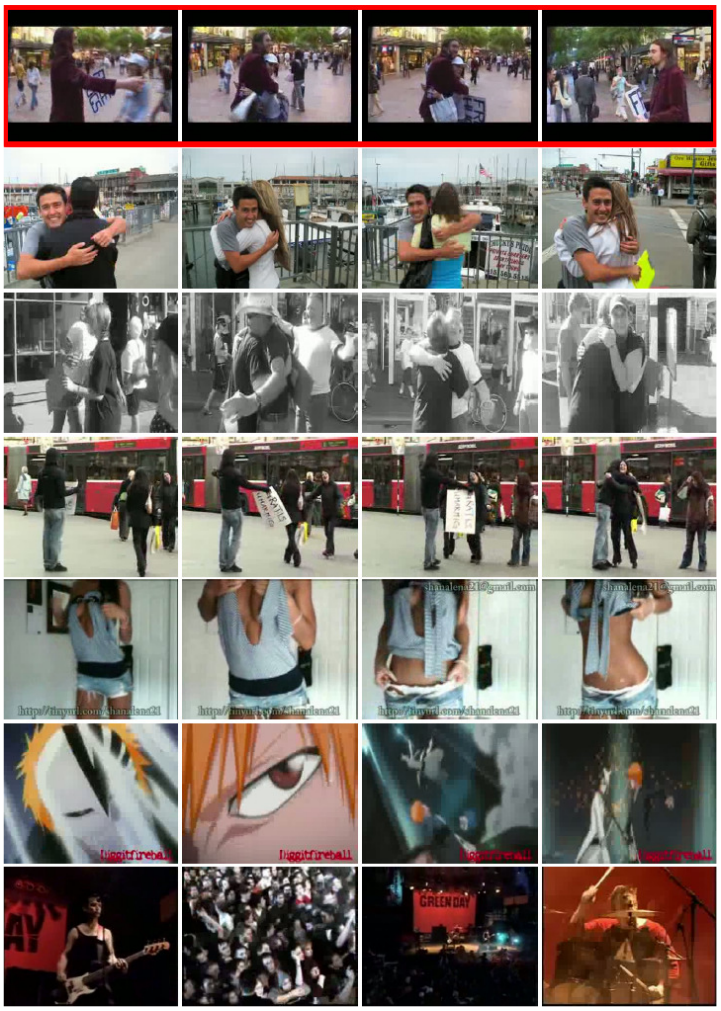

Figure 1. This figure illustrates the complexity of spam. Each line presents a few selected frames of a single video. The topmost frames (red outline) are from the original video, which is related to the Free Hugs Campaign. The next three videos (one per line) are legitimate answers to the original. The bottom three lines are spam videos.

2. RELATED WORKS

The identification of pollution in online video social networks is a new topic with very few published works. After extensive research, we could only find (Benevenuto et al. 2009) and (Langbehn et al., 2010). The first tries to identify non-cooperative users and addresses both spamming and ballot stuffing (which they call "promoting"), by analyzing parameters like tags, user profile, the user posting behavior and the user social relations. The second extends (Benevenuto et al., 2009) by proposing multi-view classification approaches to reduce the labeling cost and uses a Lazy Associative Classification (LAC) classifier. Neither work use the video content itself for classification, instead, they rely on metadata and access logs. By contrast, our scheme mainly relies on the visual content.

Some works have been proposed to detect objectionable content in visual documents (images and videos). From those, the vast majority is concerned with nudity, pornography or graphical violence. The vast majority of pornographic detectors in images or videos is based on the detection of exposed skin (Kelly et al., 2008) and seriously suffers from false positives of face close-ups, sport scenes or other innocent skin exposures. The available literature on violence detectors for content-filtering tends to concentrate on feature films and to give the soundtrack a very special attention (e.g. (Giannakopoulos et al., 2006)). That kind of specialization

warrants good performances, but makes the adaptation of those techniques difficult to the chaotic nature of social networks.

Recently, however, much attention has been devoted to less constrained approaches, using general-purpose features and classifiers. From those general approaches, one of most successful is the visual dictionary of local features.

The acceptance of local features as a broad technique of image description was an important watershed in the history of image understanding. Local features, like the popular SIFT descriptors (Lowe, 2004), allow excellent discriminating power and great robustness to geometric and photometric transformations. If they were initially available only for static images, nowadays there exist local features that take into account the spatiotemporal nature of video, one of the most popular being STIP (Laptev, 2005).

The discriminating power of local descriptors is extremely advantageous when matching objects in scenes, or retrieving specific target documents. However, when considering high-level semantic categories, it quickly becomes an obstacle, since the ability to generalize becomes then essential. A solution to that problem is to quantize the description spaces by using codebooks of local descriptors, in a technique sometimes named visual dictionary. The visual dictionary is nothing more than a representation which splits the descriptor space into multiple regions, usually by employing non-supervised learning techniques, like clustering. Each region becomes then a visual word and is included in a dictionary of visual words. The idea is that different regions of the description space will become associated to different semantic concepts, for example, parts of the human body, corners of furniture, vegetation, clear sky, clouds, features of buildings, etc. The technique has been employed successfully on several works for retrieval and classification of visual documents (Sivic and Zisserman, 2003).

In addition to moderating the discriminating power of descriptors, the dictionaries allow adapting to visual documents techniques formerly available only to textual data. Among those borrowings, one of the most successful has been the technique of bags of words (which considers textual documents simply as sets of words, ignoring any inherent structure). The equivalent in the CBIR universe has been called bag-of-visual-words, bag-of-features or bag-of-visual-features, sometimes abbreviated as BoVF. It greatly simplifies document description, which becomes a histogram of the visual words it contains. The introduction of this technique had a huge impact on content-based retrieval and classification of visual documents (Yang et al., 2007). The BoVF model also opens the opportunity to employ the Latent Semantic Analysis (LSA) (Landauer et al., 1998).

LSA was initially intended for large corpora of text, but using the metaphor of the visual words has allowed employing it for visual documents. It has been applied to image tasks, achieving good results (e.g., (Caicedo et al., 2010) and (Yanai and Barnard, 2010)). It is nothing more than an operation of change of basis in the document description in order to make more explicit some latent associations between them. Using the information provided by the bags of words, we create an occurrence matrix (telling which word occurs in which document, and usually applying some frequency normalization). LSA will then apply Singular Value Decomposition (SVD) to project this data in a new space of topics. Usually the dimensionality of the topic space will be reduced, by discarding the component of low singular value. The documents can then be described by their histogram of topics instead of words.

3. CONTENT FILTERING USING VISUAL INFORMATION

In order to identify the occurrence of spam videos in an answer thread, it is necessary establishing an approach that can be able to analyze the contextual location of each video. Then, we purpose and evaluate the approach presented in Figure 2.

In our strategy there is a Preprocessing and a Feature Extraction steps. Those, respectively, executes the preparation of the input video to the visual information extraction and the acquisition of the visual features properly. The Feature Extraction step output bag-of-visual-features representing the visual information present in the input video.

After these initial steps, the LSA is used to reduce the number of columns while preserving the similarity structure among rows. The LSA converts the bag-of-visual-features vector to a new semantic space achieving a set of latent topics, named in this work as bag-of-topics. That new space has the function of to explicit the

intrinsic relations between two or more concepts. The bag-of-topics represents the importance of each latent topic in each video.

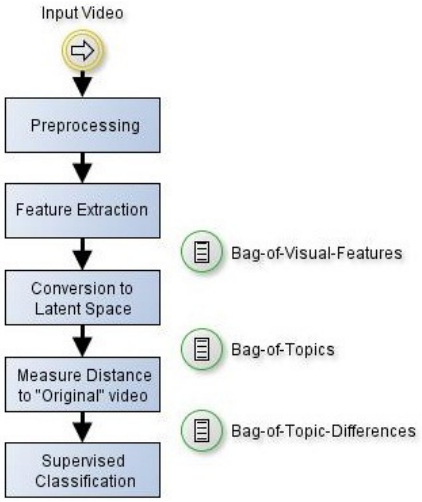

Figure 2: The proposed architecture, based on bag-of-topic-differences.

In a nutshell, while the BoVF-model takes the appearances at face value, the LSA-model looks for hidden patterns that may indicate that different visual words may be related or that certain combinations of visual words may make sense together. That might alleviate the problem of extreme visual diversity in legitimate and spam patterns.

Then is measured the similarity between each video and its original ones. Each video becomes represented by its distance to the original video of its thread. We use the influence of this measure to represents the contextual information.

Abstractly, the context-free video description was a vector in the semantic space. The context-aware description is now the vector difference between the video and its original one. This allows us to take the different contexts (the different threads) while keeping the classification model extremely simple (only two classes, spam and legitimate, for the entire dataset). We named this approach as *bag-of-topic-differences*.

The proceeded operation was very simply, but has the fundamental importance in the performance of our methodology. This step is necessary because each video thread can be associated to a different semantic concept. In the context-free features vector the landmark to values of this vector is the same to all threads. In other words, videos of different threads that occasionally have similar values will be treated as been the same thing. This situation is totally possible considering that a spam video is classified in this way only because is not related to those ones in its actual thread.

Also, is very common a video to be associated to different threads. The same video that is classified as normal video in one thread can be labeled as spam video in other. We had faced this problem generating a context-aware features vector. In context-aware we consider the original video of each thread as the landmark to all video of its thread. With this even though the same video be associated to different threads, labeled as normal in one and spam in another, the values of the features vector will be different because was considered as landmark the original video of each thread.

Figure 3 shows the visual contrast between the similarity measure in context-free feature-vectors based on bag-of-visual-features or bag-of-topics (a) and the context-aware bag-of-topic-differences (b). Two threads are represented, respectively by the blue circles and red triangles, with the landmark of the thread marked respectively A and B. The video C was proposed as an answer on both threads, but its distance to the thread landmarks suggests that it is legitimate on thread A (blue circles) and spam on thread B (red triangles).

In Fig. 3(a) the similarity between elements is measured based on distance to a common landmark for all threads. And the video in discussion (C) presents equal distance in different threads. But as it was labeled with different category will induces an error in classifier. Already in Fig. 3(b), the similarity is measured based on distance of each video to the original video of its thread. In this case, the same video present distinct distance value because the landmarks of two threads are different.

Figure 3: Visual scheme contrasting the context-free feature-vectors based on bag-of-visual-features or bag-of-topics (a) and the context-aware bag-of-topic-differences (b). Two threads are represented, respectively by the blue circles and red triangles, with the heads of the thread marked respectively A and B. The video C was proposed as an answer on both threads, but its distance to the thread heads suggests that it is legitimate on A (blue circles) and spam on B (red triangles).

4. EXPERIMENTAL RESULTS

Given the novelty of this application, it is unsurprising that no standard database is available for evaluation purposes. Therefore, we have constructed two datasets to evaluate our methodology.

The first one is a well-controlled dataset. We define 5 semantic themes (Animals, Events, Food, Personalities and Sports). And, for each one generates 4 textual searches in Youtube. Each search was about a different subtheme (e.g.: in Food we search videos of recipes with Apple, Asparagus, Broccoli and Chocolate). The first 20 retrieved videos were manually annotated. Then, we have constructed one synthetic thread of videos-response to each subtheme, containing 10 legitimate videos, 10 spam videos and 1 original video in each. The first legitimate video returned in search was annotated as original video in its thread. The spam videos were random added using the videos annotated as legitimate in others subthemes, but not used as legitimate in its thread. In resume, we creates 20 synthetic threads containing 20 videos equal distributed in both classes, legitimate and spam.

Our second dataset is an in-the-wild dataset. We have collected 8182 videos from 84 threads, chosen at random from the Most Responded Videos list generated by YouTube. We have selected the videos in the most responded list, because they form long threads, often with a lot of spam. The manual inspection determined 3420 to be legitimate and 4678 to be spam. We have used a subsample of the original dataset, randomly selecting (besides 84 originals) 1000 legitimate and 1000 spam. Deciding which videos are spam is sometimes hard, even for humans: we have considered them as spam when their visual contents did not match the subject of the thread. In case of doubt, we adopted the policy of (Benevenuto et al. 2009) and marked the video as legitimate.

The experimental design was a classical 5-fold cross-validation. In the well-controlled dataset each fold contained 320 videos for training and 80 for testing. We generate an unfavorable scenario, when the each fold was about a specific theme and any thread about this theme was added in training set. This scenario was very important to demonstrate the generalization power of the purposed approach. Is not necessary know any video about a specific theme in the training set to achieve good results in the classification task.

In the in-the-wild dataset, were used 1600 videos for training and 400 for testing in each fold. The numbers reported are the average of the folds.

We present the results with the evaluation of the proposed approach and some baselines. Baselines were performing only in the well-controlled dataset. As baselines were performed experiments with traditional bag-of-visual-features, traditional LSA (bag-of-topics) and a context-aware BoVF (bag-of-visual-differences), considering the context location of the videos using BoVF without conversion to latent space. The experiments with bag-of-topic-differences were performed using both datasets. We use the spatiotemporal (STIP) visual features. Figure 4 show the ROC graphics considering the mean of True Positive Rates (TPR) and False Positive Rates (FPR) achieved in experiments.

In all experiments, the results indicate that the context information is critical to identify spam using a two-class classification model. The visual characteristics of the video allows filtering some of the spam: this is interesting and show that, at least in our sample, videos used as spam tended to share some visual characteristics. However, even the worst context-aware experiment (filled data points) was similar than the best context-free experiment.

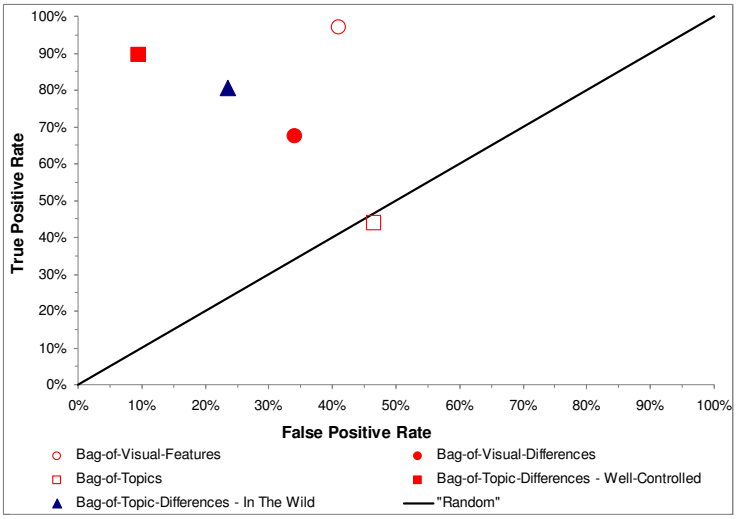

Figure 4: Experimental results with spatiotemporal feature (STIP). The sweet spot is the upper left corner. The data points represent different choices: circles × squares are bag-of-visual-features and bag-of-topics; empty × filled are context-free and context-aware experiments. All experiments considering well-controlled dataset. And, blue triangle represents context-aware topics (bag-of-topic-differences) using in-the-wild dataset.

The results indicate the great power of generalization of the approach and that the hypothesis of that the difference between a video to your video reference (original video in its thread), regardless of the semantic concept to which they belong, is able to group videos that have visual content with same semantics. Our experiments achieved good results in both datasets, considering the compromise between TPR and FPR.

Using the well-controlled dataset we can manage the mainly difficulties, the large variety of visual content in legitimate elements and the low number of one kind of class in most threads. With this dataset we achieved good results. Table 1 shows the TPR and FPR to experiments using bags-of-topic-differences with this database. As can be observed in Table 1, in four of five available themes of the well-controlled dataset the achieved results outperform 90% (TPR) and are under of 13% (FPR).

Table 2 shows the mean TPR and FPR to experiments using bags-of-topic-differences with both datasets. The achieved TPR to in-the-wild dataset demonstrates that methodology has good ability to identify spam elements. But, the high value in FPR exposes a weakness in working with a real dataset. The occurrence of few elements of one class in the thread is the main reason of the difficulty to construct a model to represents both classes.

Table 1. True Positive Rate (TPR) and False Positive Rate (FPR) for each fold using *bag-of-topics-differences* approach on *well-controlled* dataset.

Class	TPR	FPR
Animals	**0.93**	**0.00**
Food	**0.90**	**0.05**
Events	**0.95**	**0.10**
Personalities	0.70	0.20
Sports	**1.00**	**0.13**

Table 2. True Positive Rate (TPR) and False Positive Rate (FPR) using *bag-of-topics-differences* approach on both datasets.

Dataset	TPR	FPR
Well-controlled	**0.90**	**0.09**
In-the-Wild	0.81	0.24

5. CONCLUSION

Removing content automatically is only possible when false positive rates are very low, because removing a legitimate answer is much more problematic than accepting a spurious one. At this stage, our technique, used in isolation, does not allow such low rates and thus cannot be used to forcefully remove content from the social network. That does not mean, however, that the technique is of no practical value. An interesting strategy may be employed to make it feasible: combining it with manual inspection of the suspect videos, in order to remove only those which are indeed deemed as illegitimate.

The structure of experiments guarantees that the good results achieved of our approach is not dependent of the restrict dataset or a special kind of videos. There are a high visual content variation in the proposed themes and subthemes. Combined with the low quality of the results of traditional approaches, we can infer that the contextual location added by the measure of distance to a specific landmark is the responsible to improve results.

Of course, our current approach explores only the visual information contained in the video, and thus is only the lower bound on what could be obtained adding other evidences, such as those provided by the soundtrack, metadata, social network statistics, etc. Currently we are investigating the best to incorporate our visual classifier in a system taking into account all those evidences.

REFERENCES

F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida and M. Gonçalves, "Detecting Spammers and Content Promoters in Online Video Social Networks", In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 620-627, 2009.

R. Benmokhtar and B. Huet. "Multi-level Fusion for Semantic Video Content Indexing and Retrieval". In Adaptive Multimedial Retrieval: Retrieval, User, and Semantics, Lecture Notes In Computer Science, Vol. 4918. Springer-Verlag, 160-169, 2007.

J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. Gonzalez. "Combining visual features and text data for medical image retrieval using latent semantic kernels". In ACM MIR'10. ACM, New York, NY, USA, 359-366, 2010.

T. Deselaers, L. Pimenidis and H. Ney. "Bag-of-Visual-Words Models for Adult Image Classification and Filtering", In: International Conference on Pattern Recognition, pp. 1-4, 2008.

T. Giannakopoulos, D. I. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence Content Classification Using Audio features", In: Hellenic Artificial Intelligence Conference SETN-06, LNAI 3955, pp. 502-507, 2006.

W. Kelly, A. Donnellan, D. Molloy. "Screening for Objectionable Images: A Review of Skin Detection Techniques", In: International Machine Vision and Image Processing Conference, pp. 151-158, 2008.

T. Landauer, P. Foltz and D. Laham. "Introduction to Latent Semantic Analysis". Discourse Processes, 25, 259-284, 1998.

I. Laptev. "On Space-Time Interest Points", In: International Journal of Computer Vision, vol 64, number 2/3, p.107-123, 2005.

H.R. Langbehn, S.M.R. Ricci, M.A. Gonçalves, J.M. Almeida, G.L. Pappa, and F. Benevenuto, "A Multi-view Approach for Detecting Non-Cooperative Users in Online Video Sharing Systems", JIDM, pp.313-328, 2010.

D. G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints", In: International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

J. Sivic and A. Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos", In: IEEE International Conference on Computer Vision, pp. 1470-1477, 2003.

A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. "Content-Based Image Retrieval at the End of the Early Years". IEEE Trans. Pattern Anal. Mach. Intell. 22, 12, 1349-1380, 2000.

X. Wei and C. Ngo. "Fusing semantics, observability, reliability and diversity of concept detectors for video search". In ACM International Conference on Multimedia. ACM, New York, NY, USA, 81-90, 2008.

K. Yanai and K. Barnard. "Region-based automatic web image selection". In ACM MIR'2010. ACM, New York, NY, USA, 305-312, 2010.

J. Yang, Y. Jiang, A. G. Hauptmann, and C. Ngo. "Evaluating bag-of-visual-words representations in scene classification". In International Workshop on Multimedia Information Retrieval. ACM, New York, NY, USA, 197-206, 2007.

Y.J. Zhang. "Semantic-based Visual Information Retrieval". Idea Group Inc., Hershey, PA, USA, 2007.