Abstract— In this work, we investigate how much content-based visual information analysis can aid in filtering spam videos on video sharing social networks. That is a very challenging task, not only because of the high-level semantic concepts involved, but also because the diverse nature of social networks prevents the use of constrained a priori information. In addition, spam video is, by nature, context-dependent. We propose a context-aware description, which improves detection considerably in comparison with the baseline bags-of-visual-words model, by allowing us to incorporate the context of the video into the representation. Our model is evaluated in two challenging video dataset, showing very encouraging results.

Keywords- Semantic Classification; CBVIR; SVD; LSA; Bag-of-Features; SIFT.

I. INTRODUCTION

Video-sharing social networks have experienced a huge success, absorbing dozens of hours worth of new clips uploaded every minute [1]. However, that proliferation of content and the immediacy of its broadcasting has created a demand for specialized tools, both to foster legitimate uses and to inhibit abuses.

Traditional approaches, using keywords or other textual metadata associated to the video, face several issues, which persist even after requiring some mandatory annotation when the video is posted. Finding a video can be time consuming for users [2]. Therefore, social networks provide other mechanisms to retrieve videos of interest. Top videos, for example, are identified according to a variety of criteria (number of views, number of comments, user ratings, etc.). Tools are provided to explore the social aspects of the service: favorite videos, related videos, and video-responses.

The latter are a very popular feature of some social networks, like YouTube, where the user can post a video as a comment to another. However, that interesting feature may lead to abuses, when users post unrelated “answers” to the videos in the threads. We call that behavior “spamming”. Sometimes those videos contain advertisement, but not always; often they are posted in the hope to “piggyback” in the popularity of some discussion or subject. Sometimes, the misattribution is made intentionally to elicit angry responses from other users (“trolling”, in the slang of social networks).

That definition follows the one used by Benevenuto et al. [3] and is based on the semantic content of the videos rather than their appearance. The high-level interpretation of video content becomes paramount. It also makes the context where the video appears crucial.

Approaches to detect non-collaborative behavior in a video sharing social network were first presented by Benevenuto et al. [3] and extended by Langbehn et al. [4]. Both propose to detect users with non-collaborative behavior (“spammers”), using information extracted from users’ profiles, social network relationships and frequency of posting. Although our applicative context is the same, we follow a different approach in two important aspects.

First, we are trying to characterize actions instead of users, therefore our method try to classify each answer to a thread, instead of characterizing the behavior of a user in its entirety. In fact, it has been noted in both previous works [3, 4] that one of the difficulties of classifying author behavior is that there are degrees of spamming activity, and even intense spammers often post legitimate answers as well.

The second crucial difference is our choice of features, employing the visual content of the videos. As far as we know, we were the first to approach video spam using content-based analysis [5]. We believe that in the long term, all relevant information (metadata, tags, social interactions, visual content and even the audio track) will have to be analyzed in order to better combat spam; however, here we are interested in evaluating how much visual information, in isolation, is able to accomplish. Of course, nothing prevents our findings to be extended to other types of information.

The main contributions presented in this work are:

- a scheme for detecting spam in threads of video-responses, which provides powerful generalization features: it compensates small training sets sizes, and does not require training examples from each new thread to be analyzed;
- a focus on visual content, providing a system robust to the absence, or intentional manipulation of metadata.

II. TASK OVERVIEW

Spam detection in video social networks requires abilities rarely found in existing content-based classification/retrieval techniques. Any content-based solution must face the context dependent nature of spam, its diversity, the large visual variability of legitimate, and the scarcity in training
examples for each high-level concept. Finally, the diversity of concepts makes it difficult to learn a general model.

In traditional content-based video classification, one aims at assigning the videos to predefined classes, corresponding to semantic concepts (concrete ones like cars and people, or abstract ones, like entertainment and pornography). In our spam detection task, however, the classes are context-dependent, because whether or not a video is legitimate depends on where it has been posted. A soccer video in a sports thread is legitimate; but the same soccer video in a ballet thread, is spam. What characterizes legitimacy is relatedness to the thread being discussed.

The negative definition of spam (all that fails to be legitimate) makes it extremely diverse. While all legitimate videos in a single thread at least share a thematic subject, the spam videos can, each one, be about a different theme.

Another source of variability is the enormous visual diversity found in videos, even among legitimate ones of a single thread. That is one of the consequences of the social network setting, where even a very restricted theme (e.g., “how to cook asparagus”) attracts a large array of users, posting many different-looking videos.

The visual diversity of legitimate videos contributes to the “semantic gap”, i.e., the lack of coincidence between the low-level information available to the machine (e.g. color and texture descriptors) and the high-level interpretation that a user would give to that same data [6]. We must seek to reduce that gap, since our definition of spam is based not on the visual appearance, but on the theme/subject of videos.

The dependence on context leads to the difficulty in transposing knowledge obtained in one thread to the others. If a solution is not obtained to generalize the results, the model obtained to classify legitimate and spam videos in a particular thread (e.g., “how to cook asparagus”), will be useless in all others (e.g., “horses in jump training”).

In order to respond to those challenges, we would like to have a single decision model for all spam, independently of their threads. That would allow us to boost the number of training samples (by pooling the samples from all threads available for training) and eliminate the need to know every thread in advance (since the model would be general for all spam). The difficulty, of course, is to conciliate those goals with the context-dependent nature of spam. Our work will solve that dilemma by incorporating the contextual information into the features (see Section 4).

Figure 1 illustrates the diversity of the spam phenomenon. Each line presents a few selected frames. The topmost frames are from the original video, which is related to “Miss Teen USA South Carolina 2007”. We show one video per line. The second one is a legitimate response, while the third is a spam response. All videos have visual similarities, but only the legitimate is thematically related to the original.

As some terms are used in a very specific sense in this work, we define them below:

- **original video**: in a social network, video that originates a discussion thread, receiving responses in the form of other videos;
- **video-response**: video posted as a comment to another;
- **thread**: collection of video-responses to the same original video;
- **thread landmark**: reference used to compare the visual content of all videos in a thread. In this work we always use the original video as a landmark;
- **legitimate**: video-response that presents a semantic relationship, that shares the same subject, with its original video. We consider it as collaborative content;
- **spam**: video-response that does not share the subject with the original video, but has a theme of its own. We consider it as non-collaborative content.

### III. Related Works

#### A. Underpinning Tools

Our work is based on visual information extracted from videos. Low-level descriptors, like the popular SIFT features [7], allow excellent discriminating power and great robustness to geometric and photometric transformations. SIFT is both a point of interest detector, based on differences of Gaussians and a local descriptor, based on the orientations of grayscale gradients. Using SIFT, visual content is represented by a set of scale and rotation invariant descriptors. They are reasonably invariant to viewpoint and illumination change. For a comparative survey on low-level local features the reader is referred to [8].

In order to characterize the semantic concepts presented in videos, low-level features are not enough. We need a strategy to provide generalization, the so-called mid-level features. One of the most popular approaches is the bag-of-visual-words (BoVW) model. The BoVW quantize the description space using a “visual dictionary”, which is a way to split the descriptor space into multiple regions, usually by employing non-supervised learning techniques. Each region becomes then a “visual word”. That technique has been employed successfully on several works for retrieval and classification of visual documents [9, 10].

#### B. State of the Art

There are many approaches to the problem of spam in a variety of scenarios, including spam in email, spam in search engines (link building with fake web pages, for example), spam in social networks user profiles and comments, etc.
A well-known scenario is email. The goal is identifying characteristics in network traffic or textual content that enable distinguishing legitimate from spam messages [11, 12]. As spammers become more sophisticated, e-mail spam detection faces new challenges, including image spam, which consists in the use of an attached image containing the spam message, misleading thus traditional text-based algorithms. E-mail detection is a historically important example of non-cooperative behavior, but otherwise not strongly tied to our application. The interested reader is referred, for more information on image e-mail spamming, to the recent survey of Biggio et al. [13], and for e-mail spam in general, to the work of Blanzieri and Bryl [14].

Our main interest here is spam in social networks. Occurrences of fake profiles and undesirable posts are common. Irani et al. [15], Lee et al. [16] and Stringhini et al. [17] have studied approaches to detect spam bots in Twitter, Facebook and other social networks. Jin et al. [18] present data-mining based spam detection for Facebook. Their objective is to identify the posts sent from spammers, considering elements from both textual content and user behavior.

Crane and Sornette [19] use access patterns of videos by users to classify them into three categories: quality (the normal ones), viral (videos which experience a sudden surge in popularity) and junk (spurious videos, like spam).

Benevenuto et al. [3] present for the first time an explicit approach to detect non-collaborative behavior in video sharing social networks. They use features extracted from users’ profiles, relationships and posts, and apply the well-known SVM classifier. Langbehn et al. [4] present an extension to that work, by proposing a multi-view classification approach, which employs instead a lazy association-rules classifier. They were able to reduce the size of the training set, while maintaining a similar classification performance. Both works differentiate two classes of users with non-collaborative behavior: promoters (who post video- responses in order to inflate the reputation of the original video) and spammers (who post responses in an unrelated thread in order to get attention to the response). Because promoters tend to have a more predictable behavior (posting repetitive, low content responses, for example) they achieve much higher detection rates for them than for spammers. For the latter the true / false positive rates are respectively 0.58 / 0.04 for Benevenuto et al. [3]; and 0.56 / 0.08 for Langbehn et al. [4].

IV. OUR PROPOSAL

Here we try to answer the challenges of spam detection, i.e., its context-dependent nature, the large visual diversity of legitimate elements, and the difficulty of establishing a model for legitimate videos in a specific thread due to the small number of training samples. Our idea is to delegate part of the responsibility to the features, by creating a context-aware description. The goal is to obtain the generalization needed in order to create a single classification model for all threads.

In a context-free feature vector (e.g., the traditional bags-of-visual-words), visual appearance is coded as a vector in the feature space, without relation to the context of the video, which is provided in our case, by the landmark video. In other words, videos with similar visual characteristics will be near in the feature space, and videos with difference appearances will be far apart, without any regard to the threads where they have been posted. That poses a serious problem for the classifier, because the choice between spam and legitimate must be made relative to the context.

In order to solve that, we propose a context-aware description approach. Like Columbus egg, the proposition is deceptively simple: take the vector difference between the video and its landmark. That puts all videos into a new feature space, where videos close to their landmark will be close to the origin, while videos far from their landmark will be far from the origin. We call that new representation as bags-of-differences.

The context-aware description allows us to take the different threads (the different themes) while keeping the classification model extremely simple (only two classes, spam and legitimate, for the entire dataset). That simplified classification model allow us to pool the examples from all threads in the training phase, alleviating, thus, the problem of training samples scarcity.

Our proposal also enables the reuse of knowledge about legitimate and spam elements in different threads. Since the representation is made relative to the landmarks, the classifier essentially learns which regions in feature space are “close enough” to the landmark in order to be considered legitimate. Once that model is learned, completely new threads, unknown during the training phase, can be contrasted with the model.

Figure 2 contrasts the context-free representation (bag-of-visual-words) (2.a) and the context-aware representation (our approach) (2.b). Two threads are represented by the blue circles and red triangles, with their landmarks (original videos) marked, respectively, as A and B. The video marked as C was posted as a response on both threads, but its distance to landmarks suggests that it is legitimate on A (blue circles) and spam on B (red triangles). Making the representation relative to the landmark (Figure 2.b) allows us to incorporate in the feature space geometry the notion that legitimate videos should be close to their landmarks, allowing for a general classification model for all threads. That general model would be much more complex in the geometry of feature space shown in Figure 2.a.
Although the “differences” representation allows us to incorporate the context in the feature, we face the problem that the original feature space (based on visual words) is extremely anisotropic, making difficult to commensurate distances to landmarks across different regions in the space.

One way to solve that problem is to rework the geometry of the space as a whole, trying to make it more isotropic. A good candidate to achieve that is the Singular Value Decomposition (SVD) [20], which projects the data into a new space (of potentially lower dimensionality), where the dimensions have all the same scale and are uncorrelated.

One attractive thing about the SVD, is that it can be also interpreted as a Latent Semantic Analysis (LSA) [21] operation, allowing us to understand the new space as being composed of “visual topics” (linear mixtures of related visual words) instead of pure appearances. In a nutshell, while the traditional bags-of-words model takes the appearances at face value, the topics model looks for hidden patterns indicating that different visual words are related, or that certain combinations of words make sense together. That coordinated view of the features alleviates the problem of extreme visual diversity in legitimate and spam patterns.

In Figure 3, it is presented a summarized scheme of how the proposed context-aware description approach, named bag-of-topic-differences, is computed. First, we extract the low-level features (in our case SIFT features, but others could be employed), and use them to create a mid-level bag-of-visual-words representation. Then, we proceed with the feature space normalization, using the SVD. Finally, we measure the distance of each video to its landmark, making thus the description relative to the video context.

Each step may be tailored for the specific application in hand: the details of how to compute the bags-of-words or how many dimensions to keep in the SVD are not essential parts of the methodology. We give a detailed account of how we have chosen those steps in the next section.

V. RESULTS AND DISCUSSION

Given the novelty of our application, it is unsurprising that no standard dataset was available for evaluation purposes. Therefore, we have constructed two datasets to evaluate our methodology. The first one, a synthetic dataset built from searches in YouTube, was named _controlled_, and aimed at the precise evaluation of the method response to context-awareness and reuse of learned knowledge. The second, a dataset sampled from real threads of YouTube, was named _in-the-wild_, and aimed at confirming the performance in a real-world scenario.

The general methodology explained in the previous section was implemented as follows. For low-level features, we use SIFT (Lowe’s original implementation [7]). To create the bags-of-visual-words, first we pre-select at random [22] a visual dictionary with 5000 SIFT features from the dataset, which are used as visual words. Then, the bags are created as vectors of frequencies of those visual words, where a SIFT feature in a video is assigned to the closest visual word in the dictionary. The bags, which start as counts of the number of visual words present in the image, are pre-normalized using TF-IDF before SVD is applied. For the SVD, we keep all dimensions corresponding to non-zero singular values.

The vector distance to the landmarks is computed by the absolute value of the difference of each dimension.

In the experiments below we have employed an SVM classifier with linear kernel. The implementation was LIBSVM [23]. The regularization parameter of SVM (C in LIBSVM) was set to 1.0.

A. Controlled dataset

We performed a set of experiments using a dataset, where we can better control the factors to consider when interpreting the results. Those experiments allow us to evaluate the proposed context-aware approach, observing the generalization power of the learned model. In order to evaluate the proposal in a more realistic setting, we have also performed the experiments in the next section.

To build that controlled dataset, we have defined 5 main themes (Animals, Events, Food, Personalities and Sports), each with 4 subthemes, which were artificially created by performing keyword searches in YouTube. Each search was about a different concept, related to the main theme, like: “funny cats playing with a box”, for the Animals theme and...
“blocks in a basketball match”, for the Sports theme. The 20 first answers to each query were then viewed, to ensure they corresponded to the desired concept. In all subthemes we had at least 10 relevant videos. Then, we built the synthetic threads of video-answers for each subtheme. Those contain 10 legitimate videos, 10 spam videos and 1 landmark video in each. The topmost relevant video returned in each search was chosen as landmark for its thread. The spam videos were randomly added using the videos from other themes. When a video was selected to be spam, we ensured was not used as legitimate in its thread. In summary, we created 20 synthetic threads containing 1 landmark and 20 video-answers equally distributed between legitimate and spam.

The experimental design was a classical 5-fold cross-validation, where entire threads are chosen either for training, either for testing. That is done to evaluate if the classifier is able to recognize spam from new threads, from which it had no training samples. Each fold contained 320 videos (from 16 threads) for training and 80 videos (from 4 threads) for testing.

We evaluate the proposed bag-of-topic-differences (BoTD) approach and two baselines: the traditional bag-of-visual-words model (BoVW), which is oblivious to context, and a representation based on bag-of-visual-differences (BoVD), which incorporates the context information but does not normalize the feature space using the SVD (see Figure 3). We propose the latter in order to evaluate the importance of normalizing the feature space.

Figure 4 shows the ROC curves of a SVM classifier obtained using the three representations. Different points of the curve represent different biases (towards one class or the other) chosen for the classification surface, without changing its orientation. Each curve in the graph represents the mean true positive rate (TPR) of five curves obtained in the 5-fold cross validation.

The results indicate that the context information is critical in a content-based filtering using a two-class classification model, and in combination with the feature space normalization (as proposed in the BoTD model) allows for good detection performances. Interestingly, the context-free representation performs well, once we are willing to trade specificity (low false positive rate) for sensitivity (high true positive rate). But contrarily to what happens in real life, the spam videos here are chosen among a relatively small population (contrast the results of the BoVW, red curve, between Figures 4 and 5). Nevertheless, even in this first experiment, the context-free representation already losess when we are interested in high specificities (leftmost part of the curves).

B. In-the-wild dataset

The in-the-wild dataset allows us to consider some important challenges of real-world spam detection, especially in what concerns the unbounded variability of spam.

Figure 4. Experimental results on the controlled dataset. Bag-of-visual-words (BoVW) and bag-of-visual-differences (BoVD) are baselines. And, the bag-of-topic-differences (BoTD) is our context-aware proposal. The AUCs were 0.786 (BoVW), 0.674 (BoVD) and 0.955 (BoTD). In this artificial dataset, the context-free representation (BoVW) performs well for low specificities, but our proposal (BoTD) is consistently better.

Figure 5. Experimental results on the in-the-wild dataset. Bag-of-visual-words (BoVW) and bag-of-visual-differences (BoVD) are baselines. And, the bag-of-topic-differences (BoTD) is our context-aware proposal. The AUCs were 0.577 (BoVW), 0.711 (BoVD) and 0.936 (BoTD). This very diverse dataset puts on evidence the importance of the context-aware representations. The context-free features (BoVW) are no longer able to recognize spam, even for low specificities.

To construct this dataset we collected 6599 videos from 84 threads, chosen at random from the “Most Responded Videos” top list provided by YouTube. Those were selected because they tend to form long threads, often with a lot of spam. Manual inspection determined 3420 of those videos to be legitimate and 3179 to be spam. In case of doubt, we adopted the policy of Benevenuto et al. [3] and marked the video as legitimate. We did a random sub-sampling in dataset, selecting (besides the 84 original videos) 1000 legitimate and 1000 spam answers. Each original video, together with its spam and legitimate answers, was used as threads. As always, the original videos were used as landmarks.

---

1 A detailed description of the dataset can be found in www.eduardovalle.com/sups/icmc2012.pdf
interesting to note, that in this dataset where spam presents real-world diversity, the BoVW model is no longer able to recognize it, even if we allow for low specificities. The context-aware representations (BoVD, BoTD) are consistently much better than the context-free representation (BoVW).

VI. CONCLUSION

The proposed methodology does not require knowledge about any specific theme in training set. Each element is represented by its relationship to its landmark, instead of its outright visual appearance. That contributes to obtain good results even with small training sets, which are not representative of the semantic concepts involved in all threads.

Our approach explores only the visual information contained in the video, and thus it is only the lower bound on what could be obtained by adding other evidences, such as those provided by the soundtrack, metadata, social network statistics, etc. We are currently working on that direction.

ACKNOWLEDGMENT

The authors are thankful to CAPES, CNPq, FAPEMIG and FAPESP, Brazilian funding agencies, for the support to this work.

REFERENCES

[18] X. Jin, C. X. Lin, J. Luo, and J. Han, “SocialSpamGuard: A Data Mining-Based Spam Detection System for Social Media Networks”. In Int. Conf. on Very Large Data Bases, (system demo), Seattle, WA, Aug. 2011.