Covert Video Classification by Codebook Growing Pattern

Liang Du[1]
liang.du@temple.edu

Haitao Lang[2]
langht@mail.buct.edu.cn

Ying-Li Tian[3]
ytian@ccny.cuny.edu

Chiu C. Tan[1]
cctan@temple.edu

Jie Wu[1]
jiewu@temple.edu

Haibin Ling[1]
hbling@temple.edu

[1]Temple University
[2]Beijing University of Chemical Technology
[3]City University of New York

Abstract

Recent advances in visual data acquisition and Internet technologies make it convenient and popular to collect and share videos. These activities, however, also raise the issue of privacy invasion. One potential privacy threat is the unauthorized capture and/or sharing of covert videos, which are recorded without the awareness of the subject(s) in the video. Automatic classification of such videos can provide an important basis toward addressing relevant privacy issues. The task is very challenging due to the large intra-class variation and between-class similarity, since there is no limit in the content of a covert video and it may share very similar content with a regular video. The challenge brings troubles when applying existing content-based video analysis methods to covert video classification.

In this paper, we propose a novel descriptor, codebook growing pattern *(CGP), which is derived from latent Dirichlet allocation (LDA) over optical flows. Given an input video V, we first represent it with a sequence of histograms of optical flow (HOF). After that, these HOFs are fed into LDA to dynamically generate the codebook for V. The CGP descriptor is then defined as the growing codebook sizes in the LDA procedure. CGP fits naturally for covert video representation since (1) optical flows can capture the camera motion that characterizes the covert video acquisition, and (2) CGP by itself is insensitive to video content. To evaluate the proposed approach, we collected a large covert video dataset, the first such dataset to our knowledge, and tested the proposed method on the dataset. The results show clearly the effectiveness of the proposed approach in comparison with other state-of-the-art video classification algorithms.*

1. Introduction

Over last two decades have witnessed tremendous development of technologies in visual data acquisition, storage, analysis, and sharing. These technologies have been bringing great conveniences to our daily life and been affecting many research fields. On the other hand, an accompanying issue, privacy protection in dealing with visual data, has started attracting increasingly amount of academic and industry efforts [1, 2, 24, 31, 27, 32, 33]. In this paper we study the covert videos, which often relate to privacy threaten especially when captured and/or shared without authorization.

Roughly speaking, a covert video is a video such that the subject in the video was unaware of the videotaping processing. Videos taken this way mainly come from three different sources: 1) by covert surveillance system; 2) by journalist for undercover investigation; and 3) by voyeurs. Covert surveillance is originally intended to protect public and/or personal security by monitoring the behaviors of people. Nowadays, due to the fact that more and more surveillance systems are sometimes done in surreptitious manner, concerns of privacy invasion have been raised by numerous civil rights groups and privacy groups, such as American Civil Liberties Union and the legal issues related to covert video surveillance are still under debate [3, 4]. For undercover investigation, gruesome secret footages taken by covert videography have proven powerful contributions to personal/public right protection. These videos however have arguably contributed to proliferation of cases against criminalizing unauthorized entrance around the world [5]. Different from the other two sources, the intentions of voyeurs to capture covert videos are completely malicious. For example, some voyeurs spy on neighbor's home activities, others use hidden cameras to capture in public restroom, dressing room etc. Such videos often seriously

jeopardize public privacy, and when distributed on the Internet can cause worse consequences [6]. In many states and countries, such activities (capture and/or publishing of covert videos) are strictly forbidden by laws and regulations [7, 8, 9].

While covert photo classification has been recently investigated [16], classification of covert videos has never been studied to the best of our knowledge. Automatic recognition of covert videos provides an important basis toward addressing privacy issues associated with such videos. For example, if a covert video is detected when being shared publicly, an alert can be fired to trigger related operations to prevent potential leaks of privacy or security information. Classifying covert videos from regular ones, however, is a very challenging task for several reasons, including 1) large intra-class variance and between-class similarity, *e.g.*, similar actions, subjects, or scenes can be performed in both covert and regular videos; and 2) large variation in qualities of covert videos, since various video cameras can be used to secretly record videos. In addition, it is a non-trivial task to collect an effective dataset for the study due to the inherent properties of covert videos. These challenges make it hard to apply directly content-orientated video classification algorithms (*e.g.* action classification methods).

For classification of covert videos from regular videos, we propose a novel descriptor, *codebook growing pattern* (CGP), which is derived from latent Dirichlet allocation (LDA) over optical flows. Given an input video, we first calculate the histogram of optical flow, which captures frame-to-frame statistics that are characterize the videotaping process of covert videos. The histogram of optical flow are then fed into the LDA model to convert the video into a codeword string. Then, CGP is defined as the growing codebook sizes during LDA procedure. CGP naturally fits for covert video representation since (1) optical flow is capable of reflecting camera motions that characterizes covert video acquisition, and (2) CGP by itself is insensitive to video content. The CGP descriptors are compared with the χ^2 kernel for kernel-based classification.

To address the lack of covert video dataset, we have collected a large covert video dataset containing 200 covert sequences. We tested the proposed approach on the dataset together with several state-of-the-art video classification algorithms. Despite the challenge of the dataset, the experimental results show clearly the effectiveness of the proposed approach, which outperforms other algorithms.

In summary, our contribution in this paper is two-fold. On the one hand, we study a novel covert video classification problem and present a new dataset for benchmark purposes. On the other hand, we propose to use a novel video descriptor CGP and χ^2 kernel to solve this problem, which generates promising results in our experiment.

In the following of this paper, we will review related work in Section 2. In Section 3, the covert video classification method using LDA and CGP along with χ^2 kernel is proposed. We construct a benchmark of covert video classification. The introduction of this database is given in Section 4. The experimental performance and comparison of the proposed method are presented in Section 5. We conclude the paper in Section 6.

2. Related Work

As a topic in video analysis, video classification has been under active researching due to its wide range of application. Existing video classification problems are mainly defined on the video contents. Among many studies, action recognition [10] is probably the most widely studied, and spatial temporal interest points are used for characterizing different actions. Bag-of-visual-words is an important line of research due to its simplicity and robustness to noise [11, 12]. In the literature of action recognition, the studies that are most related to ours are those using topic models. For example, Wang *et al*. [13] use semi-supervised latent topic models for human action recognition. We use the latent Dirichlet allocation (LDA) for assigning codewords to each frame and a sequence of visual words are used as features for covert video classification. LDA can automatically cluster codebook without predefined number of clusters. More importantly, the stickness of the word allocation process expressed the frame to frame change properties which are crucially in covert video classification. Although there are also quickly frame-to-frame changes regular videos like ego-centric videos, we believe that there are differences between them. A sequence of the allocated visual words, termed as *i.e*. Codebook Growing Patterns (CGP) are warped in to a kernel and fed into SVM to train covert classifiers. Experiments validate our assumption.

For example, in [29] the probabilistic Latent Semantic Analysis (pLSA) is used for image scene classification. Matikaiinen *et al*. [35] propose a method to represent the spatial-temporal information using the pairwise spatial and temporal relationships. The main idea is to quantize trajectories using sequencing code map (SCM). Obviously, actions can be determined by the distribution of action-specific features, even without any temporal information as in the original bag of words framework. Unfortunately, covert videos do not have such properties. Any actions can be performed in covert videos, such as walking, running or talking. Recently, convolutional neural network has also been applied to the task of video classification [34].

Video scene classification is another video analysis task. It determines the classes of videos [14]. In this task, geometric models can be extracted for videos using information such as SFM (structure from motion) models. Nevertheless, in covert videos, any scene can appear. Also, since usually the videos are recorded in a secret condition, abruptly cov-

ering or shaking of cameras happens frequently. This makes the SFM infeasible.

Context information can be utilized to assist classification tasks [15, 17]. However, context information cannot be easily exploited in covert video classification, since similar contexts can appear in both covert and regular videos, *e.g*. talking in a meeting room can be recorded aboveboard or secretly. Some video recognition tasks related to privacy protection are related to our study, such as nudity recognition in videos [18].

Covert videos often possess large frame to frame changes due to the unstable status of video cameras. This is similar to ego-centric videos. An egocentric camera are worn on the body in order to have a natural first-person view and not needing to instrument the environment. The cameras will move with the person who wear it [19, 20, 21].

Our study is different from all previous studies mentioned above. The major difference lies in that covert video is characterized by the acquisition process, while previous studies focus on content-oriented classification. This makes content oriented solutions, *e.g*. bag-of-visual-words, hardly applicable to our task. Accordingly, we design a novel solution which uses optical flow as basic features, feeds it to LDA to generate CGP for representation, and then use a χ^2 kernel for classification.

3. Covert Video Classification

Covert videos are characterized only by the acquisition process but not the video content. This suggests that traditional content oriented video classification solutions may not be suitable for distinguishing covert videos from regular ones. Inspired by this observation, we propose a novel method by reducing the content dependency in each components. In the following, we first overview the proposed method and then detail each building blocks.

3.1. Method Overview

Our proposed covert video classification algorithm composes three major parts: the histogram of optical flow (HOF) as the low level video feature, the codebook growing pattern (CGP) as the video representation derived from the latent Dirichlet allocation (LDA), and a kernel-based classifier on the representation for classifying covert and regular videos. The flow chart of the proposed method is illustrated in Figure 1.

To summarize, for a given video V containing T frames, we first extract its HOF feature, denoted as $Y = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T) = f(V)$. The HOF feature is then fed into LDA that generates a sequence of codewords, denoted as $X = (x_1, x_2, ..., x_T) = \mathrm{LDA}(Y)$. Then from X we derive the CGP as $S = (s_1, s_2, ..., s_T) = \mathrm{CGP}(X)$, where s_t denotes the number of different codes seen so far in X. Note that, the codebooks generated by LDA are different for different videos, which makes direct inference over X meaningless. In contrast, the CGP vector S is more independent of video content and therefore suitable for describing the "covertness". Finally, a χ^2 kernel is designed for comparing CGP representations; and the kernel is combined with SVM for covert video classification. The details are given in the following subsections.

3.2. Low Level Video Representation

Histogram of optical flow (HOF) and its variants has be successfully used in many computer vision tasks, especially for videos analysis tasks, *e.g*. [22, 20, 23]. HOF is chosen for the low-level representation for covert videos.

We calculate the HOF features as following. First, the sequence of optical flow fields is computed using Lucas-Kanade algorithm. Then, the magnitude and orientation of optical flow are quantized into 3 and 8 bins respectively. In addition, in order to capture variance of optical flow vectors, their differences with the average flow with in the frame are quantized into 3 magnitudes. In total, for a video V, its HOF is a 48-dimensional feature vector, i.e., $\mathbf{y} = f(V) = (y_1, ..., y_{48})^\top$. We use this representation to capture the motion feature of video frames.

3.3. Codebook Growing Pattern from Latent Dirichlet Allocation

The HOF representation captures only short range motion information between consecutive video frames. Based on the representation, we further analyze the long range video patterns for discriminative features. For this goal, Latent Dirichlet Allocation (LDA) is conducted on HOF sequences. The benefit of using LDA is two-fold: 1) It does not need predefined number of latent topics or clusters like other topic models (*e.g*. pLSA). In covert video classification, due to the large variances of video content, it is hard to give a number of clusters in advance; 2) As shown in (1), the prior probabilities favor allocation of a frame to clusters having large numbers of frames. This "clustering effect" or "stickness" can capture the frame to frame changes of videos, which are important in discriminating covert videos. 3) The dynamic clustering procedure of LDA provides a means to investigate the code length pattern, which is insensitive to video content and can be used for covert video classification.

3.3.1 Latent Dirichlet Allocation (LDA).

Dirichlet distribution is the foundation for understanding LDA. It is a distribution over possible parameter vectors of the multinomial distribution. In fact, it can be seen as a distribution over distribution. A Dirichlet distribution is

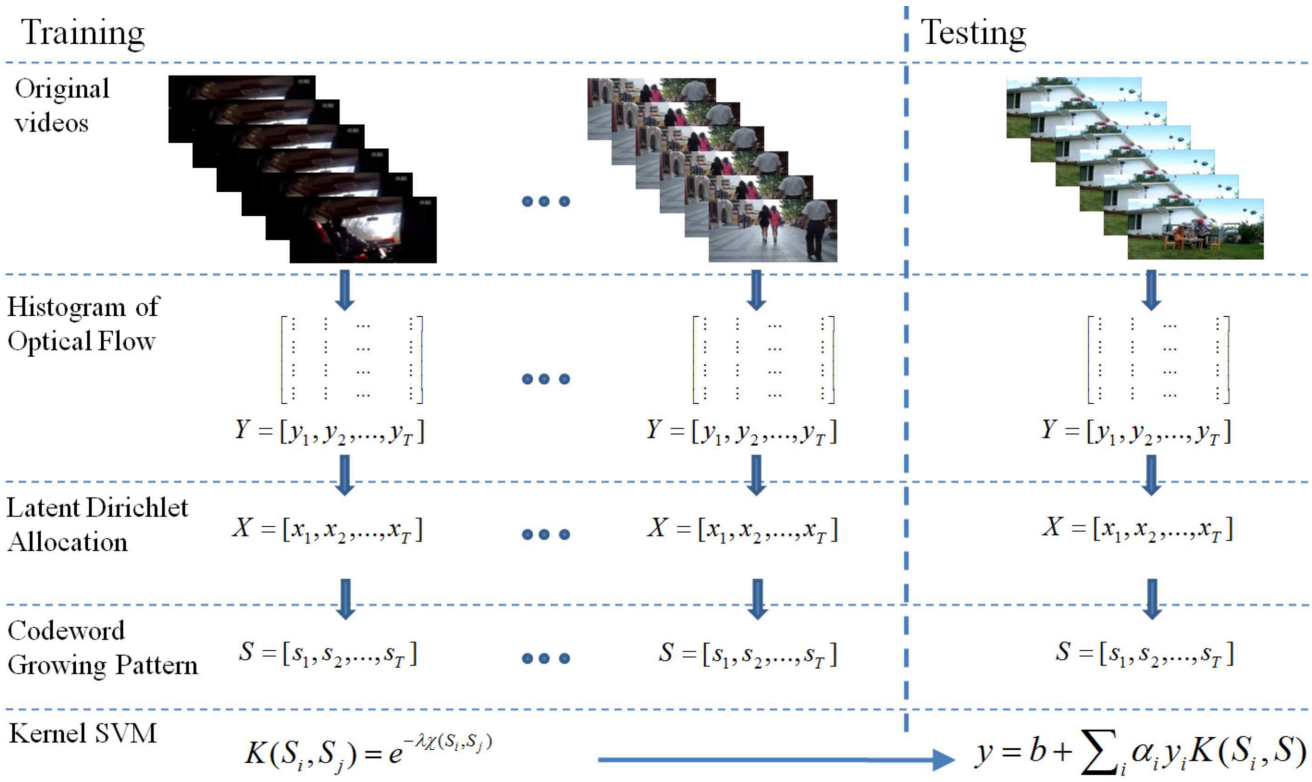

Figure 1. Framework of the proposed method

denoted as

$$G \sim \mathrm{DP}(\alpha, G_0)$$

and

$$X_n|G \sim G$$

for $n = 1, ..., N$ which means X_n is i.i.d. given G.

Marginalizing over G introduces dependencies between $X_1, X_2, \ldots, X_n$:

$$P(X_1, ..., X_n) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG.$$

Assume we view these variables in a specific order, and are interested in the behavior of X_n given the previous $n-1$. observations

$$P(X_n|X_{1:n-1}) = \\ (\frac{\alpha}{\alpha+i-1})G_0 + (\frac{1}{\alpha+i-1})\sum_{j=1}^{i-1}\delta_{X_j},\ i = 1, ..., n \quad (1)$$

Equation (1) can be understood using a Chinese Restaurant Process (CRP). Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant. From the above conditional probability distribution, we can see that a customer is more likely to sit at a table if there are already many people sitting there. However, with probability proportional to α, the customer will sit at a new table. We can rewrite (1) as

$$X_n|X_{1:n-1} = \begin{cases} X_i, & \text{with prob. } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0, & \text{with prob. } \frac{\alpha}{n-1+\alpha} \end{cases} \quad (2)$$

This is also known as the "clustering effect".

In this work, we use this model to discover the changes of video frames over time. Each video frame is represented by a feature vector $y_i, i = 1, ..., \infty$, which are observed. Here, we use histogram of oriented optical flow in each frame as our feature. The cluster X_i of y_i determined by (1) the current prior over cluster labels,*i.e.* $X_{1:i-1}$ and (2) the likelihood of the observed feature.

We use [25] for inference (2). This algorithm relies on factorizing the DP prior as a product of a prior on the partition of subjects into clusters and independent priors on the parameters within each cluster. Adding subjects one at a time, we allocate subjects to the cluster that maximizes the conditional posterior probability given their data and the allocation of previous subjects, while also updating the posterior distribution of the cluster-specific parameters.

In our task, HOF is used as the input for LDA, the process is denoted as $X = (x_1, x_2, ..., x_T) = \mathrm{LDA}(Y)$.

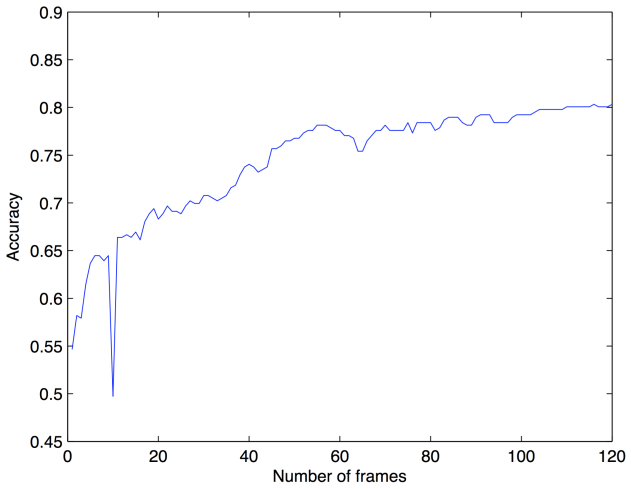

Figure 2. Average CGP for covert and regular videos.

3.3.2 Codebook Growing Pattern (CGP).

The codeword sequence X could not be used directly for comparing two video sequences. This is because LDA generates different codebooks for different videos. However, the process of dynamically introducing new clusters in LDA encodes discriminative information related to covert videos. Intuitively, covert videos often have instable motion patterns that can be captured in their optical flows. In comparison, regular videos, without any constraint, have richer patterns in their optical flows. Consequently, a covert video tends to have less clusters than a regular video in the LDA analysis of HOF.

Based on the above intuition, we propose a novel video descriptor, named *Codebook Growing Patterns*, for covert videos. The idea is to use the number of clusters during the LDA procedure to describe a video sequence. This is equivalent, given the LDA generated code sequence $X = (x_1, x_2, ..., x_T)$, to count the number of different codes in the subset $X = (x_1, x_2, ..., x_t)$, $1 \leq t \leq T$. Specifically, we define such features as $S = (s_1, s_2, ..., s_T) = \mathrm{CGP}(X)$, such that

$$s_t = \#(\{x_1, x_2, ..., x_t\}),\ 1 \leq t \leq T, \tag{3}$$

where $\#(.)$ denotes the cardinality of a set.

To show the effectiveness of CGP, we average the CGPs from two collections of videos, one for covert videos and one for regular ones. The average CGPs versus frame number t are plotted in Figure 2. It shows clearly the difference in CGP values from the two groups, especially after 60 frames. More results are given in Section 5.

3.4. Classification using Codebook Growing Patterns

Given the CGP representation, we can compare two videos by comparing their CGPs. For classification, we design a kernel on CGPs and then combine the kernel with an SVM. Noticing that a CGP vector is non-decreasing in its elements, we propose to use the χ^2 distance to compare two CGP vectors $S_i = (s_{i,1}, s_{i,2}, ..., s_{i,T})$, $i = 1, 2$:

$$\chi^2(S_1, S_2) = \frac{1}{2} \sum_t \frac{(s_{1,t} - s_{2,t})^2}{s_{1,t} + s_{2,t}}. \tag{4}$$

The distance is then used in the χ^2 kernel as below

$$K_{\chi^2}(\chi^2(S_1, S_2)) = \exp(-\lambda \chi^2(S_1, S_2)), \tag{5}$$

where λ is a constant parameter.

One assumption in the above kernel is that two sequences have the same number of frames, which is not often true in reality. However, since covert videos depend little on video content, we can use fixed length for classification. In fact, human observers usually only need to see a short segment of a video for the determination.

4. Database

To evaluate the proposed method, we collected a COVERT video dataset, containing 200 covert videos and 200 regular videos. The covert videos are collected from video sharing websites such as Youtube [1] and Youku [2]. We use key words "hidden cameras", "covert videos" to search such videos. We also checked the textual information associated with the videos to verify they are covert videos. For regular videos, to reduce bias, we intently collected hard examples, *e.g.* videos captured by amateurs, *e.g.* family videos about hiking, outdoor and indoor activities using keywords "family happy hours","hiking". Also, ego-centric videos are also included in our database. In addition, some videos from benchmark video database (UCF Action dataset [26], Hollywood Action dataset [11] *et al.*) are also included. Only one clips are selected for each action. Some example video frames are shown in Figure 3.

For the experimental protocol, we first select 50 covert videos and 50 regular ones as validation set, which is used by all algorithms for configuration and/or parameter tuning. The rest 300 videos are used for evaluation, in the leave-one-out fashion.

5. Experiments

We compared the proposed method, denoted as CGP-LDA, with the following supervised video analysis methods and latent topical models.

[1] www.youtube.com
[2] www.youku.com

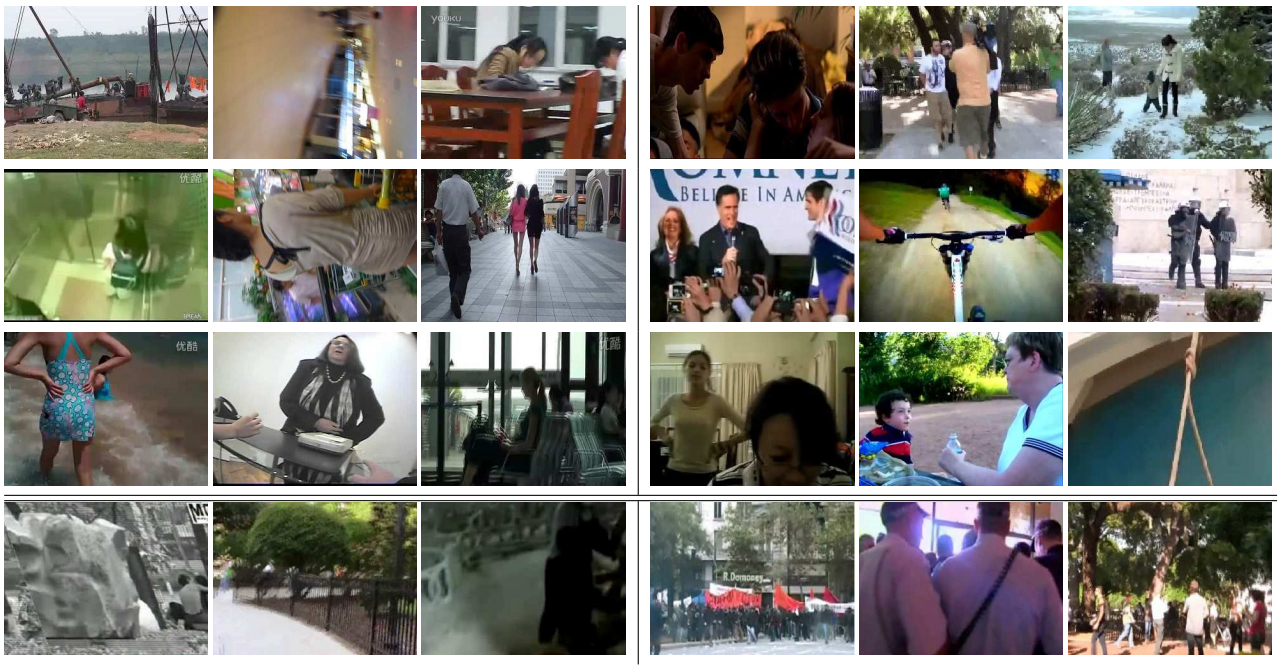

Figure 3. Example video frames from our COVERT video database. The left three columns show frames from covert videos; and the right three columns show frames from regular videos. Videos in the last row are incorrectly classified by the proposed method.

- HMM-GMM [21]: Two hidden Markov models with Gaussian mixture output are trained for both covert and regular videos. An unseen video is classified to the models with higher likelihood. This method was proposed by Spriggs et al. for first-person video temporal segmentation and action classification. Each frame is represented using the GIST feature [28]. PCA is performed on the features to reduce dimensions.

- pLSA-kmeans [29]: To test the effectiveness of LDA model in learning visual codebook, probabilistic Latent Semantic Analysis using k-means clustered codebook is tested. This experiments followed the setup of [29]. Different numbers of nearest neighbors for K-NN classification are tested.

- For the proposed CGP-LDA method, in addition to the χ^2 kernel, we also tested other kernels including the maximum difference (maxdiff) distance kernel and the Hamming distance kernel [30].

We use 100 videos, 50 positive and 50 negative, from the dataset as a validation set for parameter selection. Optimal parameters are selected and used in the rest of the database. Experiments are done in a leave-one-out framework. Classification results are reported as $Precision$, $Recall$ and $Accuracy$:

$$Precision = \frac{tp}{tp + fp},$$

$$Recall = \frac{tp}{tp + fn},$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn},$$

where tp, tn,fp and fn denote true positive, true negative, false positive and false negative respectively. In our study, covert videos are positive instances and regular videos are negative instances. The results are given in Table 1.

To further understand the property of the proposed CGP, we conducted classification experiments with different frame lengths ranging from 1 to 120. The results are shown Figure 4, which shows that larger number of frames do help improve the discriminability of CGP, though the performances tends to saturate for more than 100 frames.

The last row in Figure 3 shows some examples our method failed to classification. We observed that the wrongly classified videos are mainly due to the low video quality that leads to the unreliable low level flow features. Also, for some regular videos recorded in very unusual settings which resemble the process of covert video, *e.g.* from a long distance to record a large group of people, are incorrectly classified as covert videos.

6. Conclusion

In this paper, we studied the task of classifying covert videos, which is novel and challenging. Different from existing video analysis tasks, video content like scenes, ac-

Table 1. Evaluation Results

		Precision	Recall	Accuracy
CGP-LDA	(χ^2 kernel)	0.8041	0.8211	**0.8033**
CGP-LDA	(Maxdiff kernel)	0.8228	0.6842	0.7596
CGP-LDA	(Hamming kernel)	0.6111	**0.9842**	0.6667
pLSA	(20-NN)	0.7263	0.6244	0.6291
pLSA	(50-NN)	0.7368	0.6222	0.6291
pLSA	(100-NN)	0.7684	0.6376	0.6511
pLSA	(200-NN)	0.8368	0.5933	0.6154
HMM-GMM		**0.8529**	0.3152	0.6304

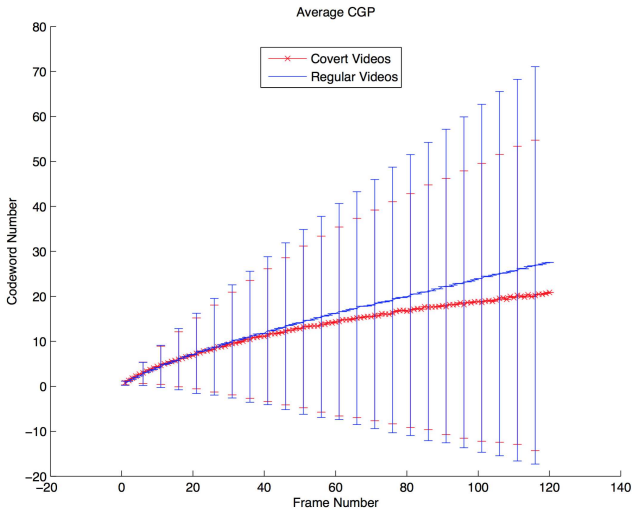

Figure 4. Video length vs. Accuracy

tions and context are unreliable. We proposed a novel descriptor named codebook growing patterns (CGP), which is derived from latent Dirichlet allocation over optical flows. The descriptor is shown to be very effective in classifying covert videos, in comparison with several state-of-the-art video classification algorithms.

Acknowledgement

This work was supported in part by NSF grants 1449860, IIS-1350521 and IIS-1218156.

References

[1] A. W. Senior, S. Pankanti, A. Hampapur, L. M. G. Brown, Y.-L. Tian, A. Ekin, J. H. Connell, C.-F. Shu, and M. Lu, "Enabling video privacy through computer vision," *IEEE Security & Privacy*, vol.3(3), pp. 50–57, 2005. 1

[2] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view," in *ICCV*, 2009. 1

[3] Web:http://www.laws179.co.nz/2011/09/covert-video-surveillance-and-covert.html "Covert video surveillance and the (c)overt erosion of the rule of law," 2011. 1

[4] L. Gray, "Police could be in breach of human rights legislation for using secret footage of hunts, say lawyers," 2010. 1

[5] A. Castrey and M. Heal, "Secret filming and the case law that subsequently arises," 2010. 1

[6] A.M. Hargrave and S.M. Livingstone, *Harm and offence in media content: a review of the evidence*, Intellect Ltd, 2009. 2

[7] The National Center for Victims of Crime, "Video voyeurism laws," . 2

[8] Daily Telegraph, "The notoriously strict privacy laws in france ensure that such intrusion into the private lives of public figures is rare," 2005. 2

[9] Web: legislation.gov.uk, "Human rights act 1998," . 2

[10] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol.43(3), pp. 16:1–16:43, Apr. 2011. 2

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008. 2, 5

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005. 2

[13] Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *PAMI*, vol.31(10), pp. 1762 –1774, oct. 2009. 2

[14] P. Gupta, S. S. Arrabolu, M. Brown, and S. Savarese, "Video scene categorization by 3D hierarchical histogram matching," in *ICCV*, 2009. 2

[15] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," in *ECCV*, 2010. 3

[16] H. Lang and H. Ling, "Covert Photo Classification by Fusing Image Features and Visual Attributes," in *IEEE Trans. on Image Processing*, 24(10):2996–3008, 2015. 2

[17] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2009. 3

[18] A.P.B. Lopes, S.E.F. de Avila, A.N.A. Peixoto, R.S. Oliveira, M. de M Coelho, and A. de A Araujo, "Nude detection in video using bag-of-visual-features," in *Computer Graphics and Image Processing*, 2009. 3

[19] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *CVPR*, 2010. 3

[20] K.M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *CVPR*, 2011. 3

[21] E.H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *CVPR Workshops*, 2009. 3, 6

[22] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, 3

[23] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, 2009. 3

[24] S. Elezovikj, H. Ling, and X. Chen, "Foreground and Scene Structure Preserved Visual Privacy Protection using Depth Information," "in *ICME*, 2013 1

[25] L. Wang and D. Dunson, "Fast bayesian inference in dirichlet process mixture models," *Computational Statistics & Data Analysis*, pp. 1–21, 2010. 4

[26] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *CVPR*, 2009. 5

[27] A. Rahmani, A. Amine, R. M. Hamou, M. E. Rahmani, and H. A. Bouarara, "Privacy Preserving Through Fireworks Algorithm Based Model for Image Perturbation in Big Data," *Int. J. Swarm. Intell. Res.*, vol. 6, pp. 41–58, 2015. 1

[28] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, pp. 145–175, May 2001. 6

[29] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA.," in *ECCV*, 2006. 2, 6

[30] R. Imre Kondor and J. D. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML*, 2002. 6

[31] M. Wilber, V. Shmatikov, and S. Belongie "Can we still avoid automatic face detection?" in *Winter Conference on Applications of Computer Vision*, 2016. 1

[32] L. Du, M. Yi, E. Blasch, and H. Ling "GARP-face: Balancing privacy protection and utility preservation in face de-identification" in *IEEE International Joint Conference on Biometrics*, 2014. 1

[33] A. Jourabloo, Z. Yin, and X. Liu, Attribute Preserved Face De-identification in *International Conference on Biometrics*, 2015. 1

[34] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale Video Classification with Convolutional Neural Networks". in *CVPR*, 2014. 2

[35] P. Matikainen, M. Hebert, and R. Sukthankar. "Representing Pairwise Spatial and Temporal Relations for Action Recognition" in *ECCV*, 2010.