SIFT-Bag Kernel for Video Event Analysis

Xi Zhou
ECE, UIUC, USA
xizhou2@uiuc.edu

Xiaodan Zhuang
ECE, UIUC, USA
xzhuang2@uiuc.edu

Shuicheng Yan
ECE, NUS, Singapore
eleyans@nus.edu.sg

Shih-Fu Chang
ECE, Columbia Univ., USA
sfchang@ee.columbia.edu

Mark Hasegawa-Johnson
ECE, UIUC, USA
jhasegaw@uiuc.edu

Thomas S. Huang
ECE, UIUC, USA
huang@ifp.uiuc.edu

ABSTRACT

In this work, we present a SIFT-Bag based generative-to-discriminative framework for addressing the problem of video event recognition in unconstrained news videos. In the generative stage, each video clip is encoded as a bag of SIFT feature vectors, the distribution of which is described by a Gaussian Mixture Models (GMM). In the discriminative stage, the SIFT-Bag Kernel is designed for characterizing the property of Kullback-Leibler divergence between the specialized GMMs of any two video clips, and then this kernel is utilized for supervised learning in two ways. On one hand, this kernel is further refined in discriminating power for centroid-based video event classification by using the Within-Class Covariance Normalization approach, which depresses the kernel components with high-variability for video clips of the same event. On the other hand, the SIFT-Bag Kernel is used in a Support Vector Machine for margin-based video event classification. Finally, the outputs from these two classifiers are fused together for final decision. The experiments on the TRECVID 2005 corpus demonstrate that the mean average precision is boosted from the best reported 38.2% in [36] to 60.4% based on our new framework.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithm, Performance, Experimentations

Keywords

SIFT-Bag, Kernel Design, Within-Class Covariation Normalization, Video Event Recognition

1. INTRODUCTION

Video based event recognition is an extremely challenging task due to all kinds of within-event variations, such as unconstrained motions, cluttered backgrounds, object occlusions, environmental illuminations and geometric deformations of objects. While there exists work attempting to detect unusual or abnormal events [37] [2] in video clips, the research on event recognition in unconstrained real-life video is still at its preliminary stage [11] [13]. We define the scope of this study as to recognize pre-defined events based on the visual cues encoded in unconstrained video, e.g. broadcast news video, as in [35] [36].

Many statistical models, e.g., Hidden Markov Model (HMM) [28], coupled HMM [3], and Dynamic Bayesian Network [27], were proposed to capture the spatial and temporal correlations of video events, and then the learnt models are utilized for pre-defined video event classification or abnormal event detection. On the other hand, appearance-based techniques were also widely used for video event detection and classification. Ke et al. [17] applied the boosting procedure for choosing the volumetric features based on optical flow representations. Niebles et al. [26] adopted the spatio-temporal interest points [8] to extract the features and other works [8, 18, 32] extracted volumetric features from salient regions [14, 18]. There also exist works that used bag-of-words model to tackle the problem of object/event recognition [39, 40]. In addition, Bagdanov et al.[41] adopted bag-of-SIFTs to detect and recognize object appearances in videos.

Most previous research on video event analysis is limited to video captured by fixed cameras in surveillance applications or greatly constrained live video. More challenging is video event recognition in unconstrained domains such as broadcast news, which contains rich information about objects, people, activities, and events [36]. For example, events in broadcast news video may involve small objects, large camera motion, and significant object occlusion, and reliable object tracking becomes very challenging.

Some recent research attempted to provide solutions for event analysis in news video. Ebadollahi et al. [10] proposed to treat each frame in a video clip as an observation and apply HMM to model the temporal patterns of event evolution in news video. Xu and Chang [35] proposed to encode a video clip as a bag of orderless descriptors obtained from mid-level semantic concept classifiers extracted from all of the constituent frames, along with the global features extracted within each video frame, and then apply the Earth Mover's Distance (EMD) [31] to integrate similarities among frames from two video clips. Multi-level temporal pyramid structure was adopted to integrate the information from different sub-clips with integer-value constrained EMD to explicitly align the sub-clips.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

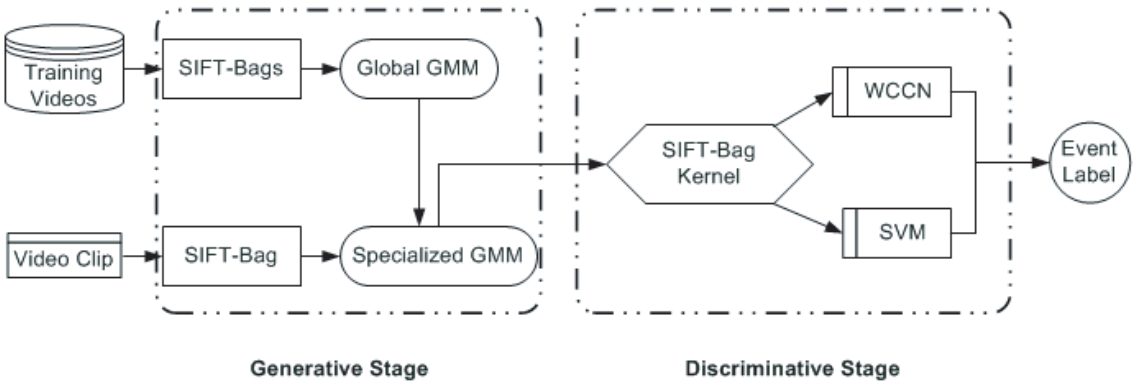

Figure 1: Overview on the SIFT-Bag based generative-to-discriminative framework for video event analysis.

Our proposed framework for video event recognition is motivated by the following observations. First, low-level global features, e.g. grid colore moments, Gabor texture histogram, and edge direction histogram, are not sufficient to characterize the image local details. Second, without the motion information, the accuracy of event recognition is still reasonably good as reported in [35]. Third, the video mismatch may exist in both spatial and temporal domains, that is, a sub-cube of one video clip may correspond to a sub-cube of another video clip belonging to the same event, but their positions and scales may be greatly different in both spatial and temporal domains. The third observation suggests video matching should be conducted based on smaller elements rather than whole frames or video clips.

In this work, we present a SIFT-Bag based generative-to-discriminative framework to address the video event recognition problem. In the generative stage, each video clip is expressed as a bag of SIFT feature vectors, motivated by the validated effectiveness of SIFT [23] in various applications. SIFT features are local and scale invariant, and hence superior over global features in expressing the local details of video frames. The SIFT-Bag representation facilitates video matching cross frames and in patch level instead of frame level. The numbers of SIFT feature vectors vary between different SIFT-Bags. Also these vectors lack correspondence and are noisy, possibly with outliers. To tackle these problems, we model the overall distribution of all SIFT feature vectors using a global Gaussian Mixture Models (GMM). Each SIFT-Bag is then represented as a specialized Gaussian Mixture Models, adapted from the learnt global GMM, base on the SIFT feature vectors within the SIFT-Bag using the Maximum a Posteriori approach.

In the discriminative stage, we design the SIFT-Bag Kernel, which is used for centroid-based and margin-based classification respectively, followed by a final fusing scheme for video event recognition. The SIFT-Bag Kernel is designed based on the so-called Super-Vector, which is derived from the upper bound of the Kullback-Leibler divergence of the specialized GMMs of any two video clips. It characterizes a simplified representation for computing the similarity between the SIFT feature vector distributions of the SIFT-Bag pair. This super-vector constitutes the SIFT-Bag Kernel in a way similar to the Gaussian kernel in common feature space. The SIFT-Bag Kernel is used for supervised learning in two ways. On one hand, to further enhance the discriminating power of the SIFT-Bag Kernel, the Within-Class Covariance Normalization (WCCN) approach is utilized to depress the kernel components with high-variability for video clips labeled as the same event, and then the refined kernel is used for similarity measurement in centroid-based video event classification. On the other hand, the SIFT-Bag Kernel is used in a Support Vector Machine [32] for margin-based video event classification. Finally, the outputs from these two classifiers are fused together for final video event recognition.

This proposed SIFT-Bag based generative-to-discriminative framework is evaluated on the large TRECVID 2005 corpus [1], and the experiments demonstrate that 1) this new framework boosts the video event recognition accuracy from the best reported 38.2% in [36] to 60.4% in term of mean average precision, and 2) SIFT-Bag representation, WCCN, and SVM all contribute to the improvement of video event recognition accuracy.

The rest of this paper is organized as follows. Section 2 gives an overview on the entire framework. The generative and discriminative stages are elaborated in Section 3 and Section 4 respectively. Section 5 provides the details on the corpus and compares experimental results of different configures of the proposed framework. The concluding remarks are given in Section 6.

2. FRAMEWORK OVERVIEW

The purpose of this work is to provide a framework, from video descriptor to classifier design, for video event recognition in unconstrained news video. Basically, this framework consists of two stages, generative stage and discriminative stage, as illustrated in Figure 1.

The generative stage includes three main parts: 1) encoding each video clip as a bag of orderless SIFT feature vectors extracted at detected salient points; 2) learning the global Gaussian Mixture Models (GMM) of the SIFT feature vector based on the SIFT feature vector ensemble extracted from all training sample video clips; and 3) deriving a specialized GMM for each video clip by using the Maximum a Posteriori approach.

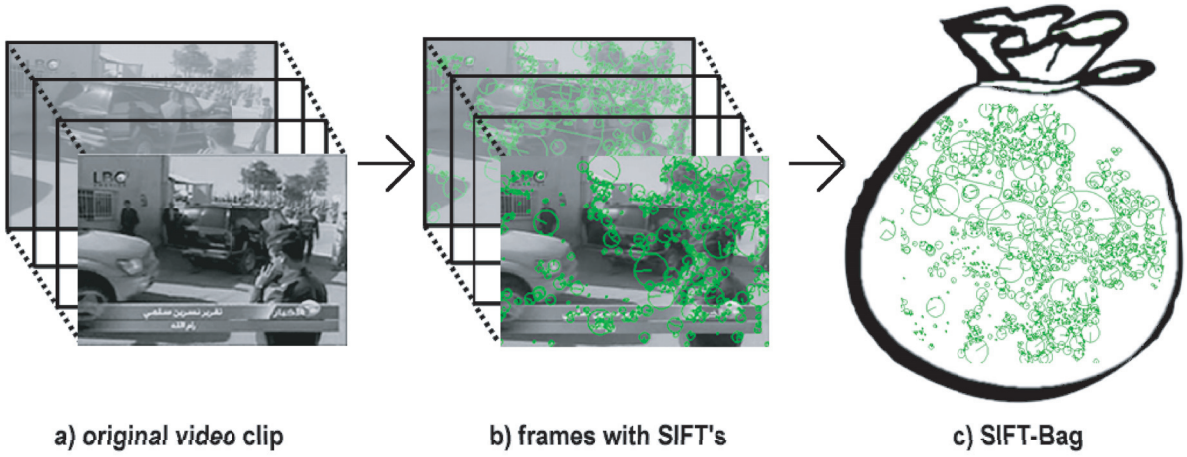

Figure 2: An illustration of SIFT-Bag extraction from video clip: a) original video frames within one video clip, b) frames with detected salient points along with the corresponding scales and dominating orientation(s), and c) SIFT-Bag. For better viewing, please see the color pdf file.

For the discriminative stage, we design the SIFT-Bag Kernel, and further use it in supervised learning for centroid-based and margin-based classification, which consequently improves the algorithmic capability in video event recognition.

More details on these two stages will be elaborated in next two sections respectively.

3. GENERATIVE STAGE

In this section, we elaborate on the generative stage of the SIFT-Bag based generative-to-discriminative framework for video event recognition.

3.1 SIFT-Bag: Video Event Descriptor

Robust feature extraction is generally critical for image and video based recognition tasks, and video event recognition requires robust features even in greater demand due to the existing complex motions, cluttered backgrounds, object occlusions, environmental illuminations, and geometric variances of objects.

Recent work of Xu and Chang [35] [36] proposed to encode a video clip as a bag of orderless descriptors obtained from mid-level semantic concept classifiers extracted from all of the constituent frames, along with the global features extracted within each video frame, and applied the Earth Mover's Distance (EMD) [21] to integrate similarities among descriptors from both video clips. These semantic concept classifiers, however, could be inaccurate, making video clip similarity measure even more challenging. Furthermore, although the semantic concept classifiers are shown effective in certain applications [35] [36], these classifiers are intrinsically application-dependent.

In this work, we propose to use bag of SIFT feature vectors as video event descriptor. Scale-Invariant Feature Transform (SIFT) [23] is a widely used algorithm to detect and describe salient local features within an image. The SIFT features are local and based on the appearance at particular interest points, and are invariant to image scale and rotation. They are also robust to changes in illumination, noise, minor changes in viewpoint, as well as occlusion. The SIFT features can be used for image matching, which is useful for object tracking and 3D scene reconstruction, and they are application-independent. In addition to these properties, they are highly distinctive, relatively easy to extract, and allow for correct object identification with low probability of mismatch. All these characteristics make SIFT feature a good candidate for video event representation.

The extraction of SIFT features consists of four major steps: (1) scale-space extrema detection, (2) keypoint localization, (3) orientation assignment, and (4) keypoint descriptor. The first step identifies potential keypoints from all locations and scales of the image. In the second step, candidate keypoints are localized to sub-pixel accuracy and eliminated if found to be unstable. The third step identifies the dominant orientation(s) for each keypoint based on the histogram of gradient in its local image patch. The assigned orientation, scale and location for each keypoint enable SIFT to construct a canonical view for the keypoint, invariant to affine transforms. The final step builds a local image descriptor for each keypoint, based upon the histogram of gradients adjusted by the dominant orientation(s). For each keypoint, we extract a SIFT feature vector (e.g., with a dimension of 128 or less).

For the video event recognition problem, the SIFT feature vectors are detected and extracted for each frame, and the SIFT feature vectors from all frames within a video clip constitute the so-called SIFT-Bag, as a video clip representation. Figure 2 illustrates the construction of a SIFT-Bag. Bag-of-Words [7] approach was widely used to transform length-variant orderless feature set into a word frequency vector of a fixed length, and then conventional machine learning algorithms can be applied based on this fixed length representation. A drawback of this Bag-of-Words model is that some useful information may be lost in the quantization process. In this work, we instead use the GMM to describe the distribution of the SIFT feature vectors within each video clip, better retaining the information in the original SIFT feature vectors. In the following two subsections, we present the details about how to obtain such a distribution model.

3.2 Global GMM for SIFT Distribution

Different from conventional Bag-of-Words framework, where each word is used as a separate entity, we estimate a GMM for the distribution of all SIFT feature vectors in the SIFT-Bag for each video clip. The reason to use a GMM for characterizing the SIFT-Bag is two-fold. First, the estimated GMM is a compact description of the underlying distribution of all SIFT feature vectors within a SIFT-Bag. Yet, with increasing number of components, the GMM can be arbitrarily accurate in describing such a distribution. The estimated GMM is less prone to noise, compared with the SIFT feature vectors themselves. Second, although explicit correspondence between SIFT feature vectors is not pursued in this framework, the Gaussian components in GMM impose an implicit multi-mode structure of the SIFT feature vector distribution in a video clip. The corresponding Gaussian components in two video clips may imply certain spatio-temporal correspondence, particularly when the GMMs for different video clips are adapted from the same global Gaussian Mixture Models as described afterwards.

Instead of separately estimating a GMM for each video clip, we estimate video clip specialized GMM by adapting from a global GMM. It is necessary and desirable, because 1) the number of the SIFT feature vectors extracted from one video clip is relatively small and insufficient for robust estimation of a GMM even in moderate scale; and 2) video clip specialized GMM adapted from the same global GMM tends to directly offer the correspondence between the Gaussian components of two GMMs.

We first estimate a global GMM using SIFT feature vectors extracted from all training video clips, regardless of their event labels. It is similar to the so-called Universal Background Model (UBM) in speech/speaker verification [30]. Then the distribution model of the SIFT feature vector for a certain video clip is adapted from the global GMM by *Maximum a Posteriori* (MAP) [20].

Here we denote $z \in \mathbb{R}^d$ as a SIFT feature vector, where $d = 64$ in this work as we use Principal Component Analysis to reduce the feature dimension from 128 to 64. The distribution of the variable z is modeled by Gaussian Mixture Models as

$$p(z;\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(z;\mu_k,\Sigma_k), \tag{1}$$

where $\Theta = \{w_1, \mu_1, \Sigma_1, \cdots\}$, w_k, μ_k and Σ_k are the weight, mean, and covariance matrix of the kth Gaussian component, respectively, and K (set as 512 in this work) is the total number of Gaussian components.

The density is a weighted linear combination of K unimodal Gaussian densities, namely,

$$\mathcal{N}(z;\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\ e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}. \tag{2}$$

We obtain a maximum likelihood parameter set of the global GMM by the conventional Expectation-Maximization (EM) approach. For computational efficiency, the covariance matrices are restricted to be diagonal [30], which proves to be effective and computationally economical.

3.3 Clip Specialized GMM by Adaptation

Generally the number of SIFT feature vectors is not enough for robustly learning the parameters of the video clip specialized GMM. On the other hand, intuitively the global GMM learnt from all training video clips may provide useful priors for the video clip specialized GMM. Therefore, it is desirable that the video clip specialized GMMs are derived in an Maximum a Posteriori way instead of an Maximum Likelihood way.

More specifically, we derive the video clip specialized GMM by adapting the mean vectors of the global GMM and retaining the mixture weights and covariance matrices. Mean vectors are adapted using MAP adaptation with conjugate priors [20], thus the parameters $\hat{\mu}_k$'s are selected to maximize

$$\begin{aligned}\ln p(\hat{\theta}, Z) &= \sum_{k=1}^{K} \ln \mathcal{N}(\hat{\mu}_k;\mu_k,\Sigma_k/r) \\ &\quad + \sum_{i=1}^{H} \ln \sum_{k=1}^{K} w_k \mathcal{N}(z_i;\hat{\mu}_k,\Sigma_k),\end{aligned} \tag{3}$$

where $\hat{\theta} = \{\hat{\mu}_1, \ldots, \hat{\mu}_K\}$ is the set of video clip specialized GMM parameters, $\Theta = \{w_1, \mu_1, \Sigma_1, \ldots\}$ are the parameters of the global GMM, and $Z = \{z_1, \ldots, z_H\}$ are the SIFT feature vectors extracted from the video clip being modeled. As shown, the conjugate prior for parameter $\hat{\mu}_k$ is itself Gaussian, $\mathcal{N}(\hat{\mu}_k;\mu_k,\Sigma_k/r)$, with a covariance matrix shrunk by a smoothing parameter r. The joint distribution function $p(\hat{\theta}, Z)$ has the same form as the likelihood function $p(Z|\hat{\theta})$, and may therefore be optimized in the same way as a likelihood function, *i.e.*, using EM with the hidden variable $Pr(k|z_i)$ as the posterior probability of the Gaussian component k for given SIFT feature vector z_i [20].

So in the E-step, we compute the posterior probability as

$$Pr(k|z_i) = \frac{w_k \mathcal{N}(z_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(z_i;\mu_j,\Sigma_j)}, \tag{4}$$

$$n_k = \sum_{i=1}^{H} Pr(k|z_i), \tag{5}$$

and then the M-step updates the mean vectors, namely,

$$E_k(Z) = \frac{1}{n_k}\sum_{i=1}^{H} Pr(k|z_i) z_i, \tag{6}$$

$$\hat{\mu}_k = \alpha_k E_k(z) + (1-\alpha_k)\mu_k, \tag{7}$$

where $\alpha_k = n_k/(n_k + r)$. MAP adaptation using conjugate priors is useful because it interpolates, smoothly, between the hyper-parameters μ_k and the maximum likelihood parameters $E_k(Z)$. If a Gaussian component has a high probabilistic count, n_k, then α_k approaches 1 and the adapted parameters emphasize the new sufficient statistics; otherwise, the adapted parameters are determined by the global model. In this work, r is adjusted, empirically, depending on the total number of SIFT feature vectors for each video clip.

4. DISCRIMINATIVE STAGE

Similar to video clip specialized GMM, we can also obtain the event specialized GMM for certain event, and then we can directly conduct video event recognition using the log likelihood ratio criteria [30], which is yet often not the best because 1) the derived event specialized model is not exactly accurate due to the assumptions in MAP adaptation for the purpose of simplicity, and 2) the learning of

global and event specialized GMMs is in a generative manner, and hence does not guarantee the optimality in discriminating power. In this section, we introduce the design of more powerful classifiers by supervised learning. More specifically, the discriminative stage of the SIFT-Bag based framework for video event recognition consists of four steps: 1) design of SIFT-Bag Kernel, 2) Within-Class Covariation Normalization (WCCN) on SIFT-Bag Kernel for centroid-based video event recognition, 3) integration of SIFT-Bag Kernel and Support Vector Machine for margin-based video event recognition, and 4) classifier fusing for final classification.

4.1 SIFT-Bag Kernel

Besides converting the data of non-fixed lengths into fixed length, designing specialized kernel is also a very popular way to apply conventional machine learning algorithms on data of variant lengths. Our work belongs to the second category. The designed kernel is derived for characterizing the KL-Divergence [24] between two video clip specialized GMMs.

Suppose we have two video clips with extracted SIFT-Bags as Z_a and Z_b. Then, from the GMM adaptation process in Equations (4-7), we can obtain two adapted GMMs for them, denoted as g_a and g_b. Consequently, each video clip is represented by a specialized GMM distribution model, and a natural similarity measure between them is the Kullback-Leibler divergence,

$$D(g_a||g_b) = \int g_a(z) log\left(\frac{g_a(z)}{g_b(z)}\right) dz. \quad (8)$$

The Kullback-Leibler divergence itself does not satisfy the conditions for a metric, but there exists an upper bound from the log-sum inequality,

$$D(g_a||g_b) \leq \sum_{k=1}^{K} w_k D(\mathcal{N}(z;\mu_k^a,\Sigma_k)||\mathcal{N}(z;\mu_k^b,\Sigma_k)), \quad (9)$$

where μ_k^a denotes the adapted mean of the kth component from SIFT-Bag Z_a, and likewise for μ_k^b. Based on the assumption that the covariance matrices are unchanged during the MAP adaptation process, the right side of the above inequality is equal to

$$d(Z_a, Z_b) = \frac{1}{2}\sum_{k=1}^{K} w_k(\mu_k^a - \mu_k^b)^T\Sigma_k^{-1}(\mu_k^a - \mu_k^b). \quad (10)$$

It is easy to prove that $d(Z_a, Z_b)^{\frac{1}{2}}$ is a metric function, and can be considered as the Euclidean distance based on the Super-Vector in another high-dimensional feature space,

$$\phi(Z_a) = [\sqrt{\frac{w_1}{2}}\Sigma_1^{-\frac{1}{2}}\mu_1^a; \cdots; \sqrt{\frac{w_K}{2}}\Sigma_K^{-\frac{1}{2}}\mu_K^a], \quad (11)$$

and then $d(Z_a, Z_b) = \|\phi(Z_a) - \phi(Z_b)\|^2$.

In this work, we use this Super-Vector to design a kernel defined as

$$k(Z_a, Z_b) = e^{-\|\phi(Z_a)-\phi(Z_b)\|^2/\delta^2} = e^{-d(Z_a,Z_b)/\delta^2}, \quad (12)$$

where δ is a constant for controlling the final similarity and we set it empirically in the experiments. Since the kernel is based on SIFT-Bag descriptor, and hence it is called SIFT-Bag Kernel hereafter.

4.2 Centroid-based Classification by WCCN

The KL-divergence is directly derived from the generative GMM and does not consider inter-class or intra-class relationships, and hence it does not necessarily provide good discriminating power. More specifically, the Super-Vector $\phi(Z_a)$ is computed directly from the video clip Z_a by adapting the global GMM, and hence is not ensured to be close to the Super-Vectors computed from video clips labeled as the same event. To further enhance the discriminating power, we propose applying Within-Class Covariance Normalization (WCCN) [4], which depresses the components with high-variability for video clips labeled as the same event.

The above kernel components are assumed in this work to be characterized by a subspace spanned by the projection matrix V. The goal of within-class covariance normalization is to identify the subspace, V, that has maximum inter-SIFT-Bag distance (maximum $\|V^T\phi(Z_i) - V^T\phi(Z_j)\|^2$) for video clip pair with the same label. Expressing this goal in the form of an optimality criterion, we find that

$$V = \arg\max_{V^TV=I} \sum_{i\neq j} ||V^T\phi(Z_i) - V^T\phi(Z_j)||^2 W_{ij}, \quad (13)$$

where W_{ij}=1 when Z_i and Z_j belong to the same event, otherwise $W_{ij} = 0$. Denote $\hat{Z} = [\phi(Z_1), \phi(Z_2), \cdots, \phi(Z_N)]$, where N is the total number of training sample video clips, then the optimal V consists of the eigenvectors corresponding to the top few largest eigenvalues of the matrix $\hat{Z}(D - W)\hat{Z}^T$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}, \forall i$.

V identifies the components in which feature similarity and label similarity are most out of sync (high label similarity corresponds to low feature similarity, and vice versa). We must de-emphasize the components $V^T\phi(Z_i)$ prior to computing the similarity.

By depressing the undesired components, the refined SIFT-Bag Kernel is defined as

$$K(Z_a, Z_b) = e^{-[\phi(Z_a)-\phi(Z_b)]^T(I-VV^T)[\phi(Z_a)-\phi(Z_b)]/\delta^2}, \quad (14)$$

where we have taken advantage of the equality $(I-VV^T)(I-VV^T) = (I - VV^T)$.

The video event recognition can be conducted directly based on the kernel similarity and the Nearest Neighbor approach. Here we use the kernel similarity between a testing video clip and the centroid of an event for similarity metric, where the centroid of an event is defined in the Super-Vector space, namely, the centroid, $\bar{Z}^s$, of the s-th event corresponds to the Super-Vector as

$$\phi(\bar{Z}^s) = \frac{1}{N^s}\sum_{i\in\pi^s} \phi(Z_i), \quad (15)$$

where Z_i is the SIFT-Bag extracted from the i-th training video clip, N^s is the number of video clips belonging to the s-th event, and π^s denotes the index set of the samples belonging to the s-th event. Then, the final video event recognition is based on normalized similarity vector as

$$C_1(Z) = [\frac{K(Z,\bar{Z}^1)}{\sum_s K(Z,\bar{Z}^s)}, \frac{K(Z,\bar{Z}^2)}{\sum_s K(Z,\bar{Z}^s)}, \cdots, \frac{K(Z,\bar{Z}^S)}{\sum_s K(Z,\bar{Z}^s)}],$$

where S (=10 in this work) is the total number of predefined events, and Z is the SIFT-Bag extracted from a test video clip.

4.3 Margin-based Classification by SVM

To enhance the discriminating power, we can also use a Support Vector Machine (SVM) [29] [12] [16] [19] [38] combined with our designed SIFT-Bag Kernel for margin-based video event classification. For a two-class problem, e.g. shooting vs. non-shooting case, the decision function for a test video clip with SIFT-Bag as Z has the following form:

$$g(Z) = \sum_i \alpha_t y_t k(Z, Z_i) - b, \tag{16}$$

where $k(Z, Z_i)$ is the value of a kernel function for the training SIFT-Bag Z_i and the test SIFT-Bag Z, y_i is the class label of Z_i ($+1$ or -1), α_t is the learnt weight of the training sample Z_i and b is the threshold parameter. The training samples with weight $\alpha_t > 0$ are called support vectors. The support vectors and their corresponding weights are learned using the standard quadratic programming optimization process or other variations. In this project, we use tools from libsvm [5] based on the multi-class SVM in our implementations.

The multi-class SVM can also output a so-called confidence vector, denoted as

$$C_2(Z) = [p_1(Z), p_2(Z), \cdots, p_S(Z)], \tag{17}$$

where $p_s(Z)$ can be roughly considered as the probability of the video clip with SIFT-Bag Z belonging to the s-th video event. Then, the classification can be conducted based on the output values in $C_2(Z)$.

4.4 Classifier Fusion

The motivations of centroid-based video event recognition and margin-based video event recognition are essentially different. Our offline experiments show that the outputs from these two classifiers are often complemental to each other, which motivates us to fuse these two classifiers to further enhance the classification capability of the whole framework. In this work, we use a simple criteria for the fusion of the outputs from these two classifiers. The vectors $C_1(Z)$ and $C_2(Z)$ both roughly measures the probabilities of a test video clip belonging to different video events, and hence we can simply average them for a more robust output as

$$C(Z) = \frac{C_1(Z) + C_2(Z)}{2}, \tag{18}$$

and then the classification can be done based on the averaged probability vector $C(Z)$.

5. EXPERIMENTS

Our experiments are conducted over the large TRECVID 2005 video corpus as in [36], and include two parts: 1) comparison of different configurations of our framework with the state-of-the-art algorithm, called **T**emporally **A**ligned **P**yramid **M**atching (TAPM) [35] [36], and 2) extensive study of algorithmic properties, i.e., SIFT-Bag visualization, evaluation by confusion matrix, and algorithmic robustness.

5.1 Corpus and Metric

As in [35], the following ten events are chosen from the LSCOM lexicon [9] [25] [36] [6]: *Car Crash*, *Demonstration Or Protest*, *Election Campaign Greeting*, *Exiting Car*, *Ground Combat*, *People Marching*, *Riot*, *Running*, *Shooting*, and *Walking*. They are chosen because these events are relatively frequent in the TRECVID data set [25] and are intuitively recognizable from visual cues. The number of video clips for each event class ranges from 54 to 877. When training the SVM, we use the video clips from the other nine events as the negative samples. We randomly choose 60% of the data for training and use the remaining 40% for testing, with the same configurations as in [35][36].

It is computationally prohibitive to compute the similarities among video clips and train multiple SVMs with cross-validation over multiple random training and testing splits. Therefore, we reported the results from the split used in [35] [36]. In the experiments, the feature extraction for all video clips costs about four hours for a 15-node computer cluster with a dual-core 2.8GHz CPU and 1G memory for each node; the global GMM training costs about one hour; the MAP adaptation for all video clips costs about 80 minutes; the WCCN and SVM step, along with the final classification, is very fast, and can be finished within few minutes.

We use non-interpolated Average Precision (AP) [33][34] as the performance metric, which is the official performance metric in TRECVID. It reflects the performance on multiple average precision values along a precision-recall curve. The effect of recall is also incorporated when AP is computed over the entire classification result set. Mean Average Precision (MAP) is defined as the mean of APs over all ten events.

5.2 Comparison With TAPM

TAPM is the state-of-the-art algorithm for video event recognition in unconstrained news video. We also got the result by Bag-of-Words quantization with SVM classification. Table 1 summarizes the comparison experimental results for different algorithms. From all these results, we can have a set of interesting observations:

1. The mean average precision is boosted from the best reported 38.2% in [36] to 60.4% based on our new framework with fusing stage.

2. For the video event of *Election Campaign Greeting*, the average precision is dramatically increased from the 13.9% to 94.8%.

3. The fusion of the two classifiers can generally further improve the average precision compared with the single classifier individually.

4. The centroid-based algorithm, namely KN+WCCN, shows to be comparable with the margin-based algorithm, namely KN+SVM.

5. Our proposed framework shows to work not as good as the TAPM algorithm for the video event of *Exiting Car*, and a possible explanation is that our framework does not explicitly pursuit temporal information, and the video event of *Exiting Car* heavily depends on the temporal contextual information.

6. The components of GMM representation (compared with Bag-of-Words quantization), SIFT-Bag Kernel design, WCCN, and SVM all contribute to the whole framework, and the best result is achieved based on the integration of them all.

Table 1: Comparison of Average Precision (%) using different algorithms. Note that: 1) TAPM-1 is the TAPM algorithm with same weights for all the three levels; 2) TAPM-2 refers to the TAPM algorithm with different weights for the three levels; 3) Hist+SVM refers to Bag-of-Words quantization with SVM classification; 4)KN+NN is the algorithm based on SIFT-Bag Kernel and Nearest Neighbor classifier; 5) KN+SVM means SIFT-Bag Kernel with SVM classification; 6) KN+WCCN refers to the centroid-based algorithm using WCCN; and 7) WCCN+SVM refers to the algorithm based on the fusion of two classifiers. The last row, referred to as Mean AP, is the mean of APs over ten events.

Event Name	TAPM-1 [36]	TAPM-2 [36]	Hist+SVM	KN+NN	KN+SVM	KN+WCCN	WCCN+SVM
Car Crash	51.1	51.0	33.0	33.5	39.7	46.5	**53.3**
Demonstration	23.6	23.6	38.2	38.3	49.3	48.5	**50.1**
Election Campaign	13.9	13.7	82.5	79.2	92.6	**94.8**	94.4
Exiting Car	**50.7**	50.1	22.1	31.5	35.2	33.9	38.1
Ground Combat	44.2	44.1	68.1	58.2	71.4	72.8	**73.4**
People Marching	25.8	25.8	70.0	67.7	75.8	76.9	**78.7**
Riot	22.7	22.9	16.9	**30.9**	24.9	25.4	27.7
Running	86.7	86.6	88.1	89.3	91.4	89.9	**91.9**
Shooting	10.4	9.9	18.0	20.0	21.9	22.7	**23.1**
Walking	52.4	52.8	52.6	59.3	73.3	66.5	**73.8**
Mean AP	38.2	38.1	49.0	50.8	57.6	57.8	**60.4**

5.3 Extensive Study

In this subsection, we present an extensive study of the SIFT-Bag based generative-to-discriminative framework in three aspects as follows.

5.3.1 SIFT-Bag Visualization

A SIFT-Bag consists of the ensemble of SIFT feature vectors extracted from a video clip. We present a visualization approach to show that by modeling the SIFT feature vector distribution of each SIFT-Bag using a GMM, we implicitly establish the correspondence between the variant numbers of SIFT feature vectors in two video clips.

First, we project the SIFT feature vector into a 2D feature space using dimensionality reduction techniques, e.g. Locality Preserving Projection [15]. All the component means of the global GMM are mapped to this 2D space. For each SIFT feature vector, its coordinates in this 2-D space are the sums of the coordinates of the component means of the global GMM, weighted by the posteriors of the components for the given SIFT feature vector.

Figure 3 shows the 2D distributions of the SIFT-Bags from three video clips, two of which belonging to the same video event of *Election Campaign Greeting*, and the other one belonging to the video event of *Running*. We can see that the SIFT feature vector distributions in the 2D space are characterized by distribution within different components, as indicated by the different colors in Figure 3. These components implicitly establish the correspondence between feature vectors in different SIFT Bags, which shows that SIFT-Bag Kernel offers the capability to match the patches from two video clips, similar in content yet different in spatial positions, scales, and temporal positions. For the video clips from the same event we can see that the feature vector distributions within the corresponding components tend to share a similar structure, while they are relatively more different for those from different events.

5.3.2 Evaluation by Confusion Matrices

Besides comparing our framework with the TAPM based on average precision in Table 1, we present more details of the performance using confusion matrices as in Figure 4.

From these confusion matrices, we observe that: 1) when evaluated by the confusion matrices, the fusion of classifiers again improves the recognition accuracy; and 2) the better the overall recognition accuracy is, the more possible the video event of *Shooting* is mis-recognized; and a possible explanation is that the event of *Shooting* is visually very similar to the event of *Ground Combat*, and cannot benefit from the improved discriminating power for most general events.

5.3.3 Algorithmic Robustness

For video event recognition, the boundaries of the video clip are often ambiguous, and also the frame rate of the video clip may vary. A good algorithm should be robust to these factors, and hence we present a set of experiments to evaluate the algorithmic robustness to these factors. In these experiments only a random portion of the frames within each video clip are used to construct the SIFT-Bag, with other aspects of the video event recognition framework unchanged.

The detailed experimental results are shown in Figure 5, with nine configurations using percentages of frames as 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% respectively. From these results, we can see that our framework is robust to the variation of boundaries and the frame rates of video clips. In particular, even when only 20% of the frames are used, our result (55.3%) still outperforms the best result (38.2%) reported in [36].

6. CONCLUSIONS

In this work, we study the challenging video event recognition problem, and propose a generative-to-discriminative framework based on SIFT-Bag representation. The SIFT-Bag representation offers the capability to implicitly match the similar contents within two video clips, different in spatial positions, scales, and temporal positions. The designed SIFT-Bag Kernel well characterizes the properties of the

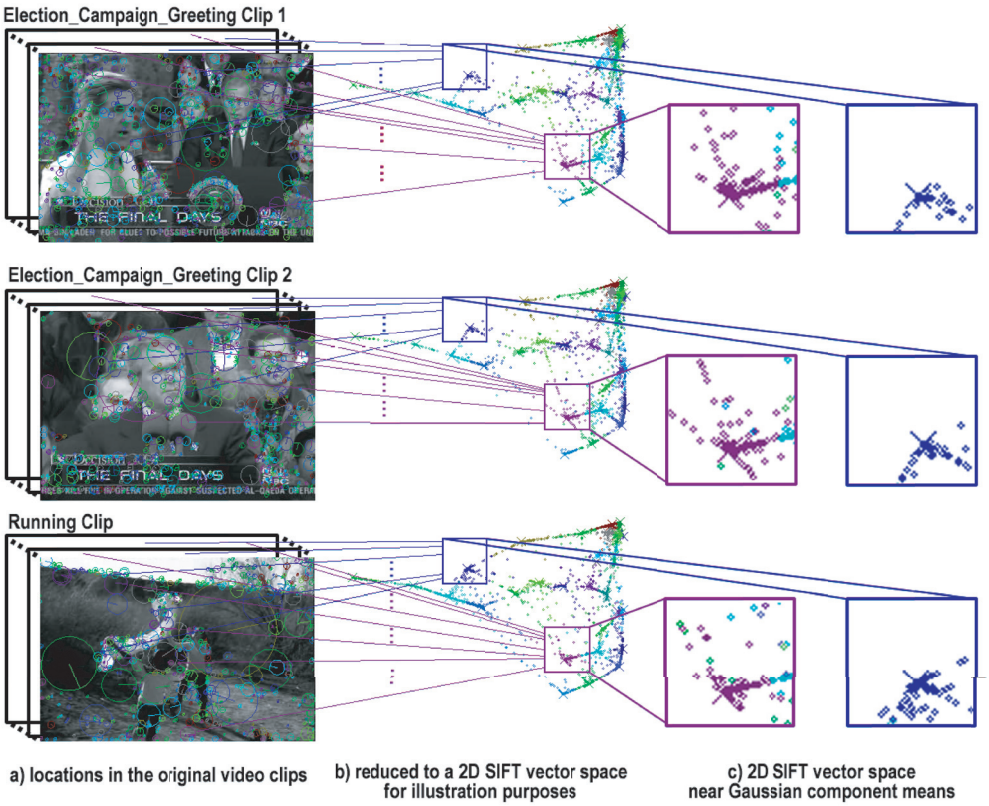

Figure 3: Visualization of SIFT-Bag in discriminating power and the capability of matching objects different in spatial positions, scales, and temporal positions. For better viewing, please see the color pdf file.

KL-divergence between two GMMs used to model the SIFT feature distributions of two video clips. The WCCN and SVM in the discriminative stage further boost video event recognition performance. The experiments shows that mean average precision of video event recognition is boosted from the best reported 38.2% in [36] to 60.4% using our proposed SIFT-Bag based framework. Our future work will focus on utilizing the motion features and spatial-temporal information to further enhance the performance of video event recognition in unconstrained news video.

7. ACKNOWLEDGEMENT

This research is supported in part by the U.S. Government VACE Program, National Science Foundation Grant IIS-0534106, and AcRF Tier 1 Grant of R-263-000-464-112, Singapore. We thank Prof. Dong Xu for providing the experiment configurations and sharing the corpus.

8. REFERENCES

[1] A. Amir *et al.*, *IBM Research TRECVID-2005 Video Retrieval System*, NIST TRECVID Workshop, 2005.

[2] O. Boiman and M. Irani, *Detecting irregularities in images and in video*, IEEE International Conference on Computer Vision, pp. 462-469, 2005.

[3] M. Brand, N. Oliver, and A. Pentland, *Coupled Hidden Markov Models for Complex Action Recognition*, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 994-999, 1997.

[4] A. Hatch and A. Stolcke, GENERALIZED LINEAR KERNELS FOR ONE-VERSUS-ALL CLASSIFICATION: APPLICATION TO SPEAKER RECOGNITION. *ICASSP*, vol. V, pp. 585-588, 2006.

[5] C. Chang and C. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. [Online]http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[6] Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts, *http://www.ee.columbia.edu/ln/dvmm/columbia374/*.

[7] L. David, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", Proceedings of ECML-98, 10th European Conference on Machine Learning: 4-15, 1998.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, *Behavior Recognition via Sparse Spatio-temporal Features*, Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65-72, 2005.

[9] DTO LSCOM Lexicon Definitions and Annotations, *http://www.ee.columbia.edu/dvmm/lscom/*.

[10] S. Ebadollahi, L. Xie, S. Chang, and J. Smith, *Visual Event Detection Using Multi-Dimensional Concept Dynamics*, IEEE International Conference on Multimedia and Expo, pp. 881-884, 2006.

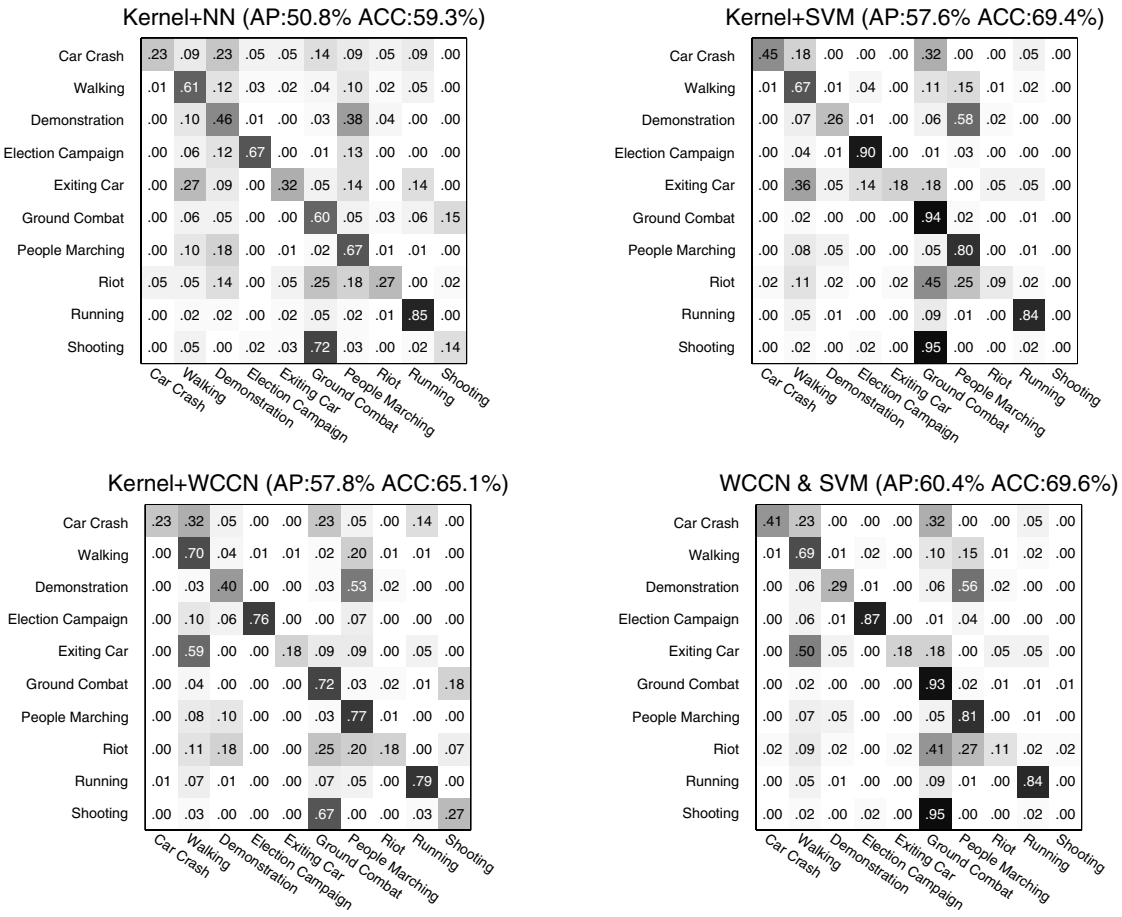

Figure 4: Comparison of confusion matrices for SIFT-Bag Kernel + Nearest Neighbor, SIFT-Bag Kernel + SVM, SIFT-Bag Kernel + WCCN, and SIFT-Bag Kernel + fusion of WCCN and SVM. Note that the first value in the title is the mean average precision, and the second value is the overall recognition accuracy. For better viewing, please see the original pdf file.

[11] A. Efros, A. Berg, G. Mori, and J. Malik, *Recognizing Action at a Distance*, Proceedings of IEEE International Conference on Computer Vision, pp.726-733, 2003.

[12] K. Grauman and T. Darrell, *The Pyramid Match Kernel: Discriminantive Classification with Sets of Image Features*, Proceedings of IEEE International Conference on Computer Vision, pp. 1458-1465, 2005.

[13] A. Hauptmann *et al.*, *Multi-Lingual Broadcast News Retrieval*, In NIST TRECVID Workshop, Gaithersburg, MD, Nov. 2006.

[14] C. Harris and M. Stephens, *A Combined Corner and Edge Detector*, Alvey Vision Confernece, 1988.

[15] X. He and P. Niyogi, *Locality Preserving Projections*, Proceedings of the Conference on Advances in Nerual Information Processing Systems, 2003.

[16] F. Jing, M. Li, H. Zhang, and B. Zhang, *An Efficient and Effective Region-based Image Retrieval Framework*, IEEE Transactions on Image Processing, vol. 13, no. 5, pp. 699-709, 2004.

[17] Y. Ke, R. Sukthankar, and M. Hebert, *Efficient Visual Event Detection Using Volumetric Features*, Proceedings of IEEE International Conference on Computer Vision, pp. 166-173, 2005.

[18] L. Laptev and T. Lindeberg, *Space-time Interest Points*, Proceedings of IEEE International Conference on Computer Vision, pp. 432-439, 2003.

[19] S. Lazebnik, C. Schmid, and J. Ponce, *Beyond Bags of Features, Spatial Pyramid Matching for Recognizing Natural Scene Categories*, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.

[20] C. Lee, C. Lin, and B. Juang, A study on speaker adaptation of the parameters of continuous density hidden Markov models. *tsap*, vol. 39, no. 4, pp. 806-814, 1991.

[21] E. Levina and P. Bickel, *The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics*, Proceedings of IEEE International Conference on Computer Vision, pp. 251-256, 2001.

[22] J. Liu *et al.*, *University of Central Florida at TRECVID 2006 High-Level Feature Extraction and Video Search*, In NIST TRECVID Workshop, Gaithersburg, MD, Nov. 2006.

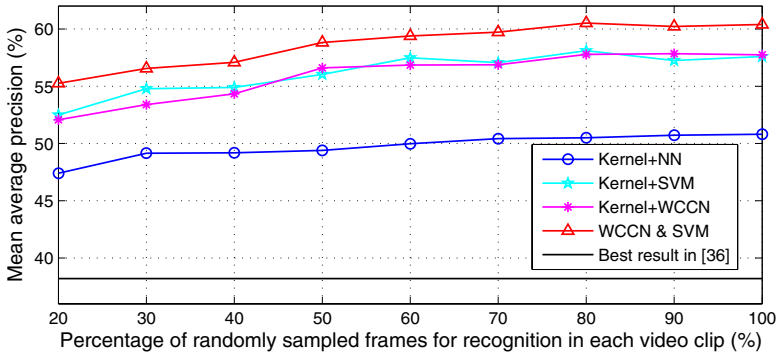

Figure 5: The comparison of mean average precisions of different algorithms using randomly sampled 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% percentages of the frames within each test video clip.

[23] D. Lowe, *Object Recognition from Local Scale-Invariant Features*, Proceedings of IEEE International Conference on Computer Vision, pp. 1150-1157, 1999.

[24] P. Moreno, P. Ho, and N. Vasconcelos, *A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications*, Proceedings of Neural Information Processing Systems, Dec. 2003.

[25] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L.Kennedy, A. Hauptmann, and J. Curtis, *Large-Scale Concept Ontology for Multimedia*, IEEE Multimedia Magazine, vol. 13, no. 3, pp.86-91, 2006.

[26] J. Niebles, H. Wang, and L. Feifei, *Unsupervised Learning of Human Action Categories Using Spatial Temporal Words*, British Machine Vision Conference, 2006.

[27] N. Oliver, B. Rosario, and A. Pentland, *A Bayesian Computer Vision System for Modeling Human Interactions*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):831-843, 2000.

[28] P. Peursum, S. Venkatesh, G. West, and H. Bui, *Object Labelling from Human Action Recognition*, Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 399-406, 2003.

[29] J. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, Advances in Large Margin Classifiers, 1999.

[30] D. Reynolds, T. Quatieri, and R. Dunn, Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[31] Y. Rubner, C. Tomasi, and L. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, International Jouranl of Computer Vision, vol. 40, no. 2, pp. 99-121, 2000.

[32] C. Schuldt, I. Laptev, and B. Caputo, *Recognizing Human Actions, A Local svm Approach*, Proceedings of IEEE International Conference on Pattern Recognition, pp. 32-36, 2004.

[33] A. Smeaton, P. Over, and W. Kraaij, *Evaluation campaigns and TRECVid*, Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321-330, 2006.

[34] TRECVID, *http://www-nlpir.nist.gov/projects/trecvid*.

[35] D. Xu and S. Chang, *Visual event recognition in news video using kernel methods with multi-level temporal alignment*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[36] D. Xu and S. Chang, *Video Event Recognition using Kernel Methods with Multi-Level Temporal Alignment*, Accepted for future publication in IEEE Transactions on Pattern Analysis and Machine Intelligence.

[37] D. Zhang, D. Perez, S. Bengio, and I. McCowan, *Semi-supervised Adapted HMMs for Unusual Event Detection*, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 611-618, 2005.

[38] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, *Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study*, International Journal of Computer Vision, vol. 73, no. 2, pp. 213-238, 2007.

[39] J. Sivic and A. Zisserman, *Video Google: a text retrieval approach to object matching in videos*, Proceedings. Ninth IEEE International Conference on Computer Vision, pp. 1470-1477, 2003.

[40] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez and T. Tuytelaars, *A Thousand Words in a Scene*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, pp. 79–86, 2007.

[41] A.D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, *Trademark matching and retrieval in sports video databases*, Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 1575–1589, 2007.