Complex Events Detection Using Data-Driven Concepts

Yang Yang and Mubarak Shah

Computer Vision Lab, University of Central Florida
{yyang,shah}@eecs.ucf.edu

Abstract. Automatic event detection in a large collection of unconstrained videos is a challenging and important task. The key issue is to describe long complex video with high level semantic descriptors, which should find the regularity of events in the same category while distinguish those from different categories. This paper proposes a novel unsupervised approach to discover data-driven concepts from multi-modality signals (audio, scene and motion) to describe high level semantics of videos. Our methods consists of three main components: we first learn the low-level features separately from three modalities. Secondly we discover the data-driven concepts based on the statistics of learned features mapped to a low dimensional space using deep belief nets (DBNs). Finally, a compact and robust sparse representation is learned to jointly model the concepts from all three modalities. Extensive experimental results on large in-the-wild dataset show that our proposed method significantly outperforms state-of-the-art methods.

1 Introduction

User uploaded videos on the internet have been growing explosively in recent years. Automatic event detection in videos is an interesting and important task with great potential for many applications, such as on-line video search and indexing, consumer content management, etc. However, it is a very challenging task to deal with large corpora of unconstrained videos with huge content variations and uncontrolled capturing conditions.(as illustrated in Fig.1).

Common approaches in event recognition rely on hand-crafted low level features such as SIFT [1], STIP [2], MFCC [3], and human-defined high level concepts [4]. The use of high level semantic concepts have been proven effective in representing complex events[5]. However, how to discover a powerful set of semantic concepts is still unclear and has not been investigated in previous works. The drawbacks of human defined concepts include: (1) it's hard to extend these concepts to a larger scale, (2) they can not handle multiple modalities, and (3) the concepts don't generalize well to new datasets.

In this paper, we propose a novel unsupervised approach to discover event concepts directly from training data in three modalities (audio, image frames and video).

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 722–735, 2012.

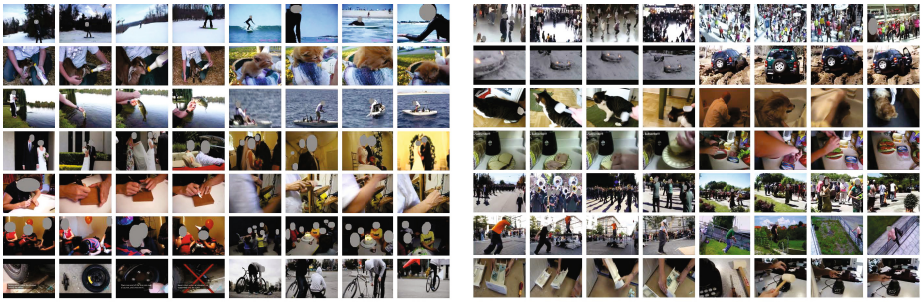

Fig. 1. Randomly selected example videos from our dataset. Each row shows frames of four videos from two categories.

We first learn low level features for each modality using Topography Independent Component Analysis (TICA), which has shown superior performance over popular hand-designed features in [6]. Then we map our low level features to more compact representations using deep belief networks (DBN)[7]. After that, our data-driven concepts are learned by clustering the training data in a low-dimensional space using vector quantization (VQ). This dimension reduction step is crucial to produce reasonable clustering results. Finally, we merge the concepts from three different modalities by learning compact sparse representations. The whole framework is shown in Fig.2.

We argue that unsupervised learning of concepts is appropriate due to two reasons. First, the disconnection between limited linguistic words and complexity of real world events makes human definition of visual concepts very hard if not impossible. We will later show that large number of learned concepts help to improve recognition accuracy significantly. Second, most of the time, insufficiency of annotated data prevents us from learning concepts in supervised manner. We present extensive evaluation of our method. The results show that our proposed approach significantly outperforms popular baselines.

The rest of the paper is organized as follows. We first review the related literature in section 2. The proposed method is presented in section 3-5 in the following order: Low-level Feature Learning, Data-driven Concept Discovery, and Event Representation Learning. Extensive experiment results, comparisons and analysis are reported in section 6. Finally, we conclude in section 7.

2 Related Work

Most of the existing works[8,5,9] investigate different hand-crafted features such as SIFT [1], STIP [2], Dollar [10] and MFCC [3]. Recently, there has been a growing interest in learning visual features using biologically-inspired networks, such as, Independent Component Analysis (ICA) [11] and Independent Subspace Analysis [12]. In [13], Le shows that using their learned 3D (Spatiotemporal) filters by ISA, the action recognition performance is comparable to other hand-designed features. In [6], TICA, another extension of ICA, was proposed for static images that achieves state-of-the-art performance on object recognition.

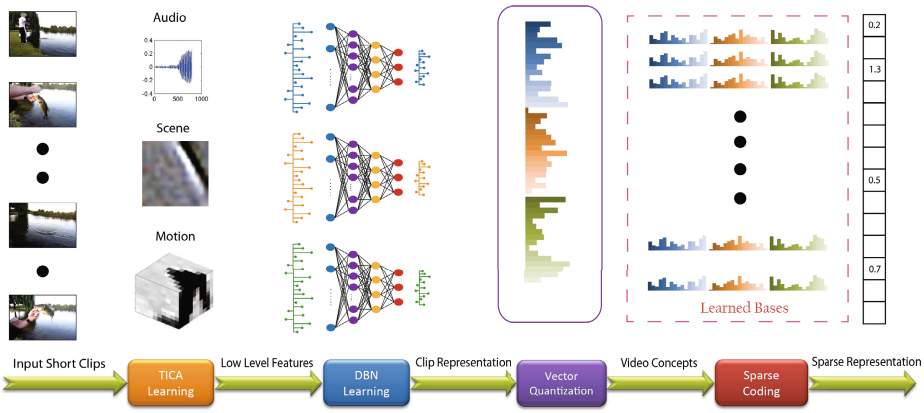

Fig. 2. Framework of the proposed method. Each video is divided into short clips. We first learn low level features for each modality using Topography Independent Component Analysis (TICA). Then we map our low level features to more compact representations using deep belief networks (DBN). After that, our data-driven concepts are learned by clustering the training data in a low-dimensional space using vector quantization (VQ). Finally, we merge the concepts from three different modalities by learning compact sparse representations.

Concept detectors provide high-level semantic representation for videos with complicated content, which can be very useful for developing powerful retrieval or filtering systems for consumer media. Lots of effort [4,14] have been devoted to building huge datasets for training concept detectors. However most of them are recorded in a well constrained conditions [15,16], which are not suitable for detecting actions in complex events.[4] provides a benchmark dataset with 25 selected concepts over a set of 1,338 consumer videos. But its concept collections are based on static images only. Audio or motion concepts are not used. Due to the large diversity of the data and insufficient training samples, concept detectors perform far below expectation. In this paper, we propose an unsupervised approach to discover concepts from three modalities using DBN, which has been proposed to solve digit recognition and achieved promising results [7]. Besides, it has been shown in [17] that DBN performs better than PCA and LLE (Locally Linear Embedding) in terms of dimension reduction.

Multiple data sources can be combined using either early fusion or late fusion strategies [18,4,5,8]. Traditional fusion methods treat each source independently[5]. We argue that it is desirable to exploit the relationships between multiple sources to achieve robust classification. In this paper, we propose sparse coding [19] to perform late fusion and empirically show the benefits of such approach.

3 Learning Low-Level Features

We use the TICA feature learning networks [6] to learn the invariant audio, scene and motion features from 1D audio signal, 2D image patches and 3D video

cuboids respectively. To make the paper self-contained we briefly descbribe TICA in the context of event recognition. For more details, please refer to [6] and [20]

We write $x^{(p)} \in \mathbb{R}^n$ as the p^{th} whitened local raw signal extracted from one modality of the video clips. For 2D image patches and 3D video cuboids, we flatten them into 1D vectors. Learning features can be viewed as learning a set of filters that map the raw signal into feature space by calculating the filter responses. TICA is a two-layered network. The first layer learns m filters $S \in \mathbb{R}^{m \times n}$ from input $x^{(p)}$ by minimizing Eqn.3. The filter responses are the activations of the first hidden layer units H. The second layer's filters V are manually fixed to pool over a small neighborhood of adjacent first layer units H, representing the subspace structure of the neurons in the first layer. More specifically, in a 2D topography, the h_k units lie on a 2D grid, with each activation of the second layer r_i pooling over a connected 3×3 block of H units through V.

In more detail, the activation of units k in the first layer is:

$$h_k\big(x^{(p)}; S\big) = S_k \cdot x^{(p)}, \tag{1}$$

where S_k is the k^{th} row of S.

The activation of unit i in the second layer is:

$$r_i\big(h_k; V\big) = \sqrt{\sum_{k=1}^{m} V_{ik} h_k^2}, \tag{2}$$

where $V \in \mathbb{R}^{m \times m}$ is a fixed matrix that encodes the topography of the hidden units H. m is the number of hidden units in the first layer.

In the filter learning process, the optimal S is learned by minimizing function:

$$\begin{aligned} S^* = \ & \arg\min_{S} \sum_{p=1}^{T} \sum_{i=1}^{m} r_i\big(x^{(p)}; S; V\big). \\ & \text{s.t.} \quad SS^T = I \end{aligned} \tag{3}$$

where T is the total number of training samples. The orthogonality constraint $SS^T = I$ provides competitiveness and ensures that the learned features are diverse. In the feature extraction process, given S^* and the new whitened local raw signal x, the activation in the second layer R will be served as the feature of x.

Considering the data we have are quite diverse and huge amount, we argue that learning good features directly from the data is very efficient. We choose TICA as our low-level building block because of its two advantages: feature robustness and less computational complexity. The pooling architecture of TICA ensures the learned features are invariant to slight location and orientation shifts, and selective to frequency, rotation and motion velocity. The filter learning is much faster than other methods such as GRBM [21] since the gradient of the objective function Eqn.3 is tractable. The feature extraction is also fast compared with sparse coding as the feature is simply computed through the matrix vector products.

4 Data-Driven Concept Discovery

Previous works[15,16] use human defined concepts for action recognition. However, this is not suitable to event recognition due to two reasons: first, defining concepts that describe the huge diversity of human actions using limited linguistic words is not practical. Second, current event datasets [22] do not have detailed annotation of action concepts for each video clip, which make it hard to train concept detectors. These two problems originally motivate us to propose data-driven concept discovery.

We assume there is only one type concept from each of the three modalities appears in a single shot clip. The idea is that we want to find a representation $Y \in \mathbb{R}^d$, which can map each clip of different modalities from raw signal space to a semantic space, where clips with similar concepts are near with each other. Considering the high diversity of our data, instead of pooling the low-level TICA features spatial-temporally, we use bag of word (BoW) histogram $Q \in \mathbb{R}^D$ by adopting vector quantization (VQ) technique using K-means soft assignment [23].

One problem is, the BoW histogram is usually long (corresponding to large cookbook) in order to capture variations of data. And k-means is well-known to be sensitive to noise in a high dimensional space especially when we apply Euclidean distance as similarity measurement. To address this, we propose to use deep belief nets (DBN) [17] to learn a lower dimensional representation for the clips from each modality. A DBN is a two-layered network, which is a stack of restricted Boltzmann machines (RBMs). The activations of the lower RBM serve as the input of the upper RBM. In each RBM, the hidden layer captures strong correlations of the units' activities in the layer below. For our highly complex event data, stacking several RBMs is an efficient way to progressively expose low-dimensional, non-linear structure. We begin by describing RBM in the case of real value input following the description in paper [17]and [24], and then we show how we use the learned clip representation to discover data-driven concepts from each modality.

DBN Learning: We start with the visible units Q in the bottom layer, which are essentially the BoW representation of each clip. A set of hidden units l are built through symmetric connection weights represented by weight matrix W. We can view the RBM as an undirected graphical model and the energy of any state in it is given by the following function:

$$\begin{aligned} E(q,l) &= -\log P(q,l) \\ &= \frac{1}{w\sigma^2}\sum_i q_i^2 - \frac{1}{\sigma^2}(\sum_i c_i q_i + \sum_j b_j l_j + \sum_{i,j} w_{ij} q_i l_j). \end{aligned} \tag{4}$$

Here, σ is the standard deviation of the Gaussian density, l_j are hidden unit variables, q_i are visible unit variables, w_{ij} is the weight connected with q_i and l_j, c_i and b_j are the bias term of visible and hidden units respectively. The learning process is to estimate w_{ij}, c_i and b_j through minimizing the energy of states drawn from the data q distribution, and raise the energy of states that are

improbable given the data. We follow [24] to use contrastive divergence learning which gives an efficient approximation to the gradient of the energy function. Further, in each iteration, we apply contrastive divergence update rule, followed by one step of gradient descent using the gradient of the regularization term.

Once finish training a layer of the network, we feed the output values of this layer as inputs of the next higher layer. Finally, after finishing training all the layers, we obtain the clip representation as the outputs of the last layer, denoted as $Y \in \mathbb{R}^d$. By doing so, We map the original features to much lower dimensional space since $D \ll d$.

Building Concepts: The low dimensional representations Y from similar clips are then grouped into concepts with a semantic meaning. In our framework, concepts are obtained from three modalities separately and each event video is represented as the occurrence frequency of each concepts from three modalities, denoted as Z.

5 Event Representation Learning

It is common that concepts of different modalities are highly correlated with each other. For example, in a birthday party event, action concept 'people dancing' is almost always co-occur with concept 'happy music' or scene concept 'crowd people', instead of 'horrible music' nor 'traffic scene'. By modeling the interaction context and inter-modality occurrence of concepts, we can removing noisy concepts and further improve the event representation. The idea is that we want to learn a set of bases which capture the co-occurrence information of concepts and the event can be represented as a linear combination of the bases. Further by imposing the sparsity on the coefficients, the noisy occurrence of irrelevant concepts will be removed.

More precisely, given N events represented in terms of concatenated concepts from three modalities, $\{Z^{(1)}, \cdots, Z^{(i)}, \cdots, Z^{(N)}\}$. We learn the basis by modeling it as a sparse coding problem [19]:

$$\phi^* = \arg\min_{a,\phi} \sum_i \left\| Z^{(i)} - \sum_j a_j^{(i)} \phi_j \right\|_2^2 + \beta \|a^{(i)}\|_1$$
$$\text{s.t.} \quad \|\phi_j\|_2 \leq 1, \quad \forall j \in \{1, 2, \cdots, s\}. \tag{5}$$

where ϕ_j is the basis vector, $a_j^{(i)}$ is the coefficient of i^{th} event associated with j^{th} basis. The first term in Eqn.5 is reconstruction error, while the second term enforces sparsity of coefficients. β is the relative weight to balance the two terms. We use sparse coding algorithm in [19] to solve this minimization problem.

After learning a set of bases ϕ, we can encode an input event $Z^{(t)}$ as sparse linear combination of the bases. The combination coefficients $a^{(t)}$ will serve as the final representation of this event, and can be obtain by solving Eqn.6.

$$\arg\min_{a^{(t)}} \left\| Z^{(t)} - \sum_j a_j^{(t)} \phi_j \right\|_2^2 + \beta \|a^{(t)}\|_1. \tag{6}$$

SVM with χ^2 kernel is used for classification [25].

6 Experiments

In this section, we will describe the dataset sand discuss several interesting observations that we had.

6.1 Datasets and Experimental Settings

We tested our approach on TRECVID 2011 event collection [22], which has 15 categories: "Boarding trick", "Flash mob", "Feeding animal", "Landing fish", "Wedding", "Woodworking project", "Birthday party", "Changing tire", "Vehicle unstuck", "Grooming animal", "Making sandwich", "Parade", "Parkour", "Repairing appliance", "Sewing project". As shown in Fig.1, it is a new set of videos characterized by a high degree of diversity in content, style, production qualities, collection devices, language, etc. The frame rate ranges from 12 to 30 fps, resolution ranges from 320×640 to 1280×2000, the time duration ranges from 30 seconds to 5 minutes.

We manually defined and annotated an action concept data set based on TRECVID event collection, which has 62 action concepts(e.g. open box, person cheering, animal approaching, wheel rotating, etc.) for approximately 9,000 videos. To the best of our knowledge, this is the largest action concepts dataset in related literatures.

In the experiment, we first resize all the videos to 480×640, and then divide each video into 4 and 10 seconds clips with 2 seconds overlap, based on our observation that the motion concepts duration varies from 2 to 10 seconds. There are approximately 300,000 clips in total. Performance was evaluated in terms of Mean Average Precision (MAP) on 15 events.

Also, we compare our low-level features with other hand-designed features on UCF YouTube action dataset, which has 11 action categories. 25-fold cross-validation is used. It is important to note that although the YouTube dataset is one of the most extensive realistic action datasets in the vision community, it is still less noisy and much simpler then the TRECVID data in terms of inner-class diversity.

6.2 Low-Level Feature Extraction

We use TICA to learn three modalities of feature representation: audio, image and video. For each modality, approximately 200,000 sampled signal/patches/video-blocks are used to train the filters and 600 filters from each modality are finally chosen as the bases for feature construction. The audio signal is extracted with sampling rate of 16 KHz. The inputs of the visual layer are 800, 20×20, $20 \times 20 \times 10$, respectively, in the three modalities. Fig. 3,4,5 shows randomly selected learned filters on 1D, 2D and 3D training examples respectively.

In order to demonstrate that our learned features outperform other classical hand designed features. We use the same bag of word framework as [26] where the code book is generated using K-means and the histogram is classified using SVM with χ^2 kernel. The code book size is set to 4,000. We compare our results on

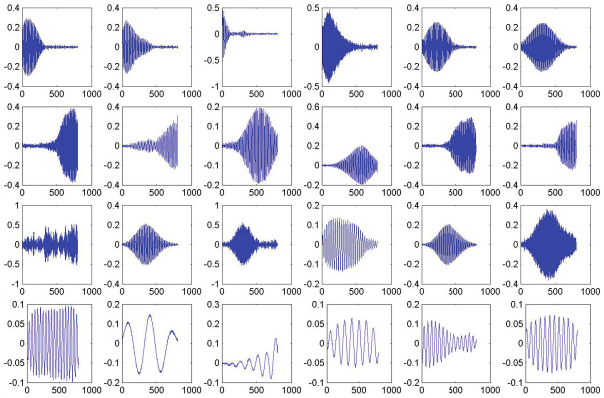

Fig. 3. 24 out of 600 audio filters learned from TRECVID event collection

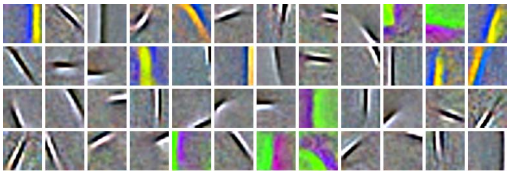

Fig. 4. 48 out of 600 image (2D) filters learned from TRECVID event collection. Since the training patches have 3 channels (RGB), the learned filters are also with 3 channels. The color information of the filters mainly captures the scene concepts, such as indoor, outdoor.

manually annotated 62 action concepts, EC and UCF 11 dataset, using MFCC [3], MBH [27], SIFT [1] and STIP [2].

Table 1 summarizes the results. Our learned features work 10% better on average in terms of recognition accuracy than all the other hand designed features, on EC and 62 concepts dataset. The performance of our 3D TICA feature is 20% higher than STIP (30.9%) motion feature on 62 action concepts dataset. The results also show that combining features from different modalities improves the overall accuracy. This suggests that features from different modalities capture complementary information and using features from different sources is necessary.

Interestingly, we notice that the learned features are not better than Motion Boundary Histogram (MBH) on UCF-11 dataset. The reason is presumably that MBH tends to overfit itself to relatively easy dataset, such as UCF-11, which contains only well-defined action with relatively simple and clean background. However, its performance plunges by a half to 44% on difficult dataset TRECVID, where our method yields the highest accuracy of 63.2%. This demonstrates the robustness of learned local features and suggests that feature discovery is important and necessary especially under uncontrolled in-the-wild condition.

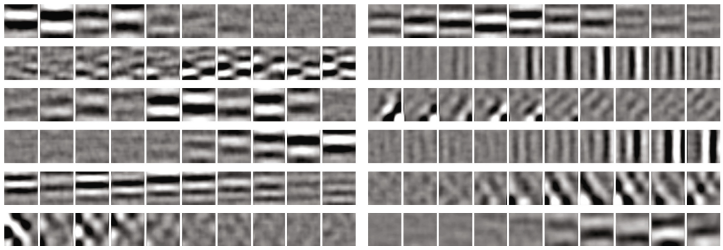

Fig. 5. 12 out of 600 spatiotemporal (3D) filters learned from TRECVID event collection. Filter size is $20 \times 20 \times 10$. Each row shows two 3D filters.

Table 1. A comparison of performance using different features and modality combinations. Our learned features outperform all the other hand designed features on the difficult TRECVID dataset; Combining features from different modalities improves the overall accuracy.

	UCF 11	TRECVID 62 concepts	TRECVID 15 Events
MFCC [3]	×	31.1	34.8
SIFT [1]	58.1	40.3	30.1
STIP [2]	57.5	30.9	41.0
MFCC+SIFT+Dollar	×	×	**51.1**
MBH [27]	**83.9**	36.0	44.0
ISA[13]	75.8	51.3	53.5
TICA 1D	×	31.7	39.7
TICA 2D	56.4	43.3	45.2
TICA 3D	74.3	**53.5**	**55.2**
TICA 2+3D	79.1	57.9	59.5
TICA 1+2+3D	×	58.1	**63.2**

6.3 Data-Driven Concept Discovery

We trained a five-layer deep belief net using RBM at each layer. The RBMs were initialized with small random weights and zero biases, and trained for 60 epochs using mini-batches size of 100. For the linear-binary RBM we used a learning rate of 0.001. We reduced the learning rate at the beginning of learning when the gradient can be large, and also at the end of learning in order to minimize fluctuations in the final weights. We also used a momentum of 0.8 to speed up the learning.

Fig.6 shows the detection results, which evaluates the performance of the clip representation of each layer: after being trained in each layer, clips are grouped into concepts based on the new representation. Then, SVM is used for classification. We first attempt to use K-means directly on the initial clip representation without any dimension reduction on the data. The event detection MAP is 39% based on concept representation. Then we adopt RBM recursively to reduce the

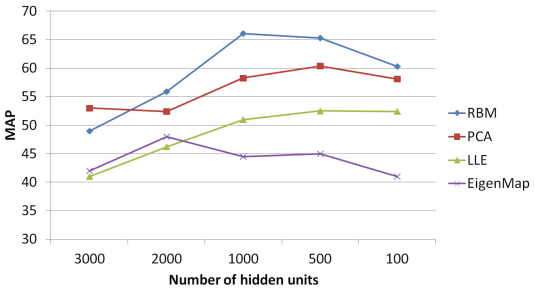

Fig. 6. A comparison event detection performance as a function of number of hidden units (dimensionality) of RBM, PCA, LLE and EigenMaps. The accuracy of directly using K-means clustering on 4000 dimension without any dimension reduction is 39%. We compare different dimension reduction techniques. RBMs achieve the best performance of 66% when the number of hidden units is 1,000. Note that the accuracy generally increases, for all techniques, while we reduce the dimension of clip representation, which suggests the necessity of dimension reduction in concept learning.

data dimension from 4,000 to 100. Fig.6 shows that when the dimension of the clip representation reduces from 4,000 to 1000, the event detection MAP increases from 39% to 66% and it reaches the highest point at 1,000 dimension. This supports our assumption that the initial representation of the clip lies in a high dimensional space where Euclidean distance can not measure the true similarity and DBN learns the regularity between the clips correctly. If we keep decreasing the dimension of the clip representation, the accuracy goes down. It means that the high dimensional data is compressed into a too concise space, some useful information maybe lost there. Further, we repeat the same experiments using other manifold learning methods such as PCA, EighenMap and LLE. Figure 5 shows the detection results. It is clear that DBN performs significantly better.

Fig.7 shows the classification results based on a different number of concepts using three modalities and their combination. The results show that motion concepts play an important role in the event detection problem. And a large number of concepts helps the recognition mainly because that larger pool of concepts captures finer level variations of actions, e.g., running action from different viewpoints. However, when the number of concepts increases further, the accuracy drops presumably due to the insufficiency of training video samples, and SVM runs into overfitting.

We observe that, the curve of audio signal (TICA 1D) peaks at 500 concepts, while that of scene (TICA 2D) and motion (TICA 3D) reach their highest performance much later. This implies that the underlying variation of audio signal is less than that of motion and 2D scene signals, which is consistent with common sense. Combined concepts (TICA 1D+2D+3D) achieve the best results since audio, scene and motion concepts capture complementary information in the videos. We also use STIP feature to run the same experiments.

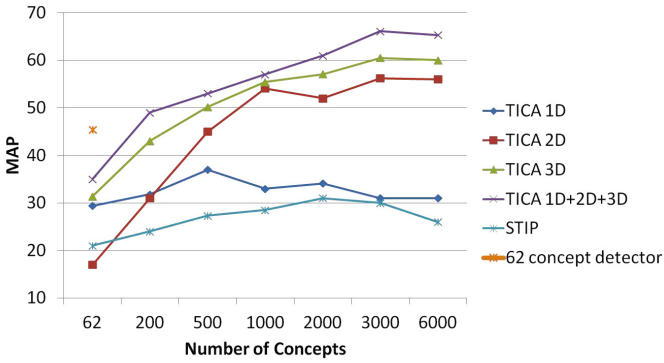

Fig. 7. A comparison of event detection performance using proposed approach employing audio, image and video features individually and jointly as a function of number of discovered concepts. We also show performance using standard STIP features. Finally, we show the importance of discovered concepts compared to manually annotated 62 action concepts. The results show that a larger number of data-driven concepts improves the detection rate and is better than human annotated concepts.

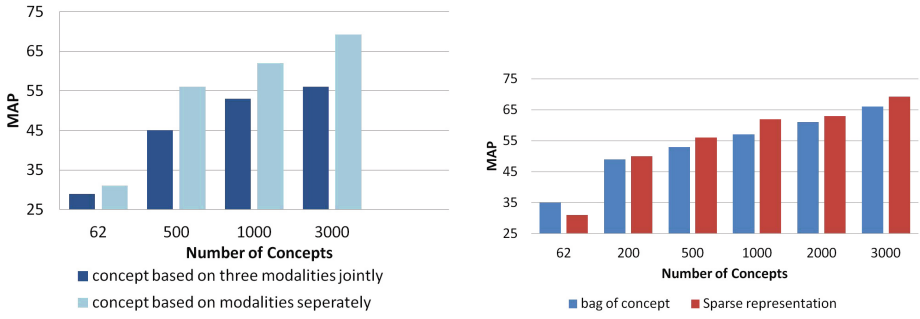

Fig. 8. (a)A comparison of event detection using different number of concepts discovered using three modalities jointly and separately. The former fuse the low level features from three modalities first as the initial clip representation, then learns concepts jointly. The later discovers the audio, scene and motion concepts separately. The results show that using concepts based on the modalities separately outperform the early fusion one. (b)A comparison of event detection using sparse representation and bag of concept representation. Sparse representation works consistently better than bag of concept representation.

The performance is significantly worse than using TICA feature. This shows that low-level features are important for discovering meaningful data-driven concepts.

Interestingly, the model using 62 human-defined concepts trained in supervised manner, outperforms its counterpart using the same number of concepts

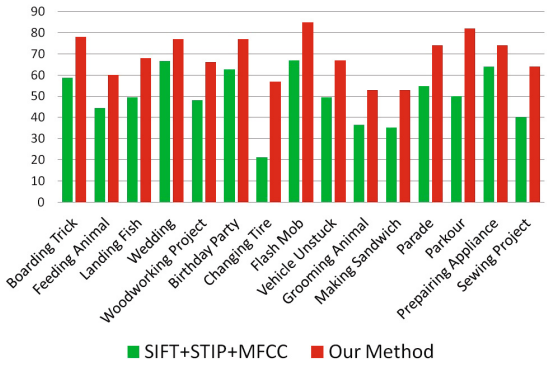

Fig. 9. Comparison of our proposed method with the combination of MFCC, SIFT and STIP features in terms of detection accuracy on each event category. Our mean average precision is 68.2%. The MAP of combination method is 51.1%.

but discovered in unsupervised way. This suggests more supervision helps when only a small number of concepts are used. However, when the number of concepts increase, data-driven concepts perform better. This shows that a large number of concepts improves the event detection.

We also compare the performance of discovered concepts using early feature fusion of three modalities to the concepts learned from three modalities separately. Fig.8a shows that discovering concepts separately is always better.

6.4 Sparse Video Representation

After concepts are discovered, each long event video can be represented in terms of concepts. Fig.8b shows the comparison of the sparse representation and the bag of concept representation, in terms of detection rate. In addition, based on the best results, the detection rate of each category compared with baseline (SIFT + MFCC + STIP) is shown in Fig.9. The MAP of our method over 15 events is **68.2%**. In comparison, the MAP of SIFT+STIP+MFCC is 51.1%.

7 Conclusion and Future Work

In this paper, we present a three step approach which learns the sparse video representation based on data-driven concepts from three modalities (audio, image and video) in an unsupervised manner. Through learning the low-level features and clip representation, high-level semantic concepts are discovered. Extensive experiments show that our method significantly outperforms the baselines using human designed features on complex in-the-wild event recognition dataset.

Acknowledgements. The research presented in this paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department

of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110 (2004)
2. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, pp. 432–439 (2003)
3. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition, Englewood Cliffs, New Jersey. Prentice-Hall Signal Processing Series (1993)
4. Loui, A.C., Luo, J., Chang, S.F., Ellis, D., Jiang, W., Kennedy, L.S., Lee, K., Yanagawa, A.: Kodak's consumer video benchmark data set: concept definition and annotation. In: Multimedia Information Retrieval, pp. 245–254 (2007)
5. Wei, X.Y., Jiang, Y.G., Ngo, C.W.: Concept-driven multi-modality fusion for video search. IEEE Trans. Circuits Syst. Video Techn. 21, 62–73 (2011)
6. Le, Q.V., Ngiam, J., Chen, Z., Chia, D., Koh, P.W., Ng, A.Y.: Tiled convolutional neural networks. In: NIPS 2010 (2010)
7. Hinton, G.E., Osindero, S., Whye Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation (2006)
8. Schindler, G., Zitnick, L., Brown, M.: Internet video category recognition. In: CVPRW 2008, pp. 1–7 (2008)
9. Wang, Z., Zhao, M., Song, Y., Kumar, S., Li, B.: Youtubecat: Learning to categorize wild web videos. In: CVPR 2010, pp. 879–886 (2010)
10. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
11. van Hateren, J.H., Ruderman, D.L.: Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex (1998)
12. Hyvärinen, A., Hoyer, P.: Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Computation (2000)
13. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011, pp. 3361–3368 (2011)
14. Liu, X., Huet, B.: Automatic concept detector refinement for large-scale video semantic annotation. In: 2010 IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 97–100 (2010)
15. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR (2009)
16. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. IEEE Computer Society (2008)
17. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science 313, 504–507 (2006)

18. Snoek, C.G.M.: Early versus late fusion in semantic video analysis. ACM Multimedia, 399–402 (2005)
19. Olshausen, B.A.: Sparse coding of time-varying natural images. In: Proc. of the Int. Conf. on Independent Component Analysis and Blind Source Separation, pp. 603–608 (2000)
20. Hyvarinen, A., Hoyer, P., Inki, M.: Topographic ica as a model of v1 receptive fields. In: IJCNN 2000, vol. 4, pp. 83–88 (2000)
21. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional Learning of Spatio-temporal Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 140–153. Springer, Heidelberg (2010)
22. Trecvid (2011), http://www-nlpir.nist.gov/projects/tv2011/tv2011.html
23. van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1271–1283 (2010)
24. Lee, H., Ekanadham, C., Ng, A.Y.: Sparse deep belief net model for visual area V2. In: Advances in Neural Information Processing Systems 20, pp. 873–880 (2008)
25. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software http://www.csie.ntu.edu.tw/~cjlin/libsvm
26. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, p. 127 (2009)
27. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR 2011, pp. 3169–3176 (2011)