Multimedia Event Detection using Visual Features

Dishan Gupta, Rahul Goutam, Amos Ng
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{ dishang, rgoutam, ajng }@cs.cmu.edu

Abstract

Learning spatial features from static images has traditionally involved approaches such as SIFT, HOG and SURF, to name a few. These approaches typically learn low-level hand-designed features which are difficult and time-consuming to extend to the video domain. Furthermore, recent research has shown that there is no universal set of hand-designed features for all datasets. Therefore, learning features directly from the dataset in an unsupervised manner has become popular. In this project, we use unsupervised feature learning to learn features directly from video data. More specifically, we use the independent subspace analysis(ISA) algorithm to learn invariant spatio-temporal features coupled with deep learning techniques like stacking and convolution to learn hierarchical representations. This approach was taken by Le et al.[14] for the task of action recognition. We apply it to the more complicated task of Multimedia Event Detection. We present our results on a subset of the TRECVID 2011 Multimedia Event Detection dataset using only visual features as well as a fusion of visual and audio features (the audio features were developed by team 2).

1 Introduction

In the past decade, the Internet has experienced an explosion of multimedia content. Therefore, multimedia content analysis has become a fundamental research issue aimed at applications such as indexing and retrieval. Since it is difficult to characterize an event, the task is much more challenging than traditional computer vision tasks such as object recognition or even action recognition. An event is an "activity-centred happening that involves people engaged in process-driven actions with other people and/or objects at a specific place and time"[9]. For example, the event "attempting a board trick" includes video clips such as skating, skiing, surfing and even finger skating. Previous research[9] has shown that, due to the complexity of the task, employing both high and low level features during classification is more useful than just one of them. For example, a video containing visual concepts such as 'man', 'animal', 'grass' and 'food' is probably about 'feeding an animal'.

It is clear that feature representation is one of the fundamental problems in multimedia event detection. Unfortunately, till recently, most of the focus in this area has been on hand-crafted features like SIFT [15], MoSIFT [26], HOG [5] and STIP [11]. The problem with such features is that they are not robust, invariant and do not work well over different domains. In recent years, unsupervised feature learning techniques like Sparse Coding [17], Deep Belief Networks [7], and Stacked AutoEncoders [3] have been used that learn to extract hierarchical feature representations from the data. Such techniques learn features directly from the data and have been shown to work well [21], providing evidence to the usefuleness of automating the feature extraction process.

In this report, we used the approach taken by [14] on the task of action recogntion to the task of multimedia event detection. More specifically, we used independent subspace analysis (ISA) algorithm [8] combined with convolution and stacking for unsupervised feature extraction. Even though the

ISA algorithm learns features from video that are robust to local translation, it is very slow to train if the dimension of the input data is large which is always the case when dealing with high-resolution video data. Therefore, we used the ISA algorithm combined with a deep learning technique called convolution. We trained the ISA on small blocks of the input first and then convolved the trained network with the rest of the input to get the entire set of features. Subsequently, for hierarchical learning (learning high-level features from low-level features) we used another deep learning technique called stacking, where we stacked a second ISA on top of the first layer.

After unsupervised feature learning, we experimented with chi-square kernel SVM and random forest classifier for the task of event classification. We evaluated our method on a subset of the dataset released for the TRECVID 2011 Multimedia Event Detection [1] task. The entire dataset consists of approximately 2000 videos with each video describing a single event. There are 15 events in total for the entire dataset. Within the given timeframe for this project, we could only run experiments on 5 of the 15 events and show the results in this report. We also present our results after fusion of both visual and audio features on the same 5 events.

2 Related work

In the past, literature related to event detection has been sparse even though problems such as video concept annotation have been widely studied. For example, [22] have integrated visual features, speech features and frequent semantic patterns of videos for annotation. In another related work, Snoek et al. [20] have worked on automatically indexing 101 semantic concepts.

Most of the existing research on event detection has focused on sports events, news events or events with repetitive patterns. Such events are generally predefined such that detecting them based on a learned set of event-specific rules or templates is feasible. [27] proposed using web-casting text and broadcast video to detect events from live sports games .

Most of the work on image classification uses hand-crafted features like SIFT [15], MoSIFT [26], STIP [11] and HOG [5] with reasonable success. Extending such features for video (3D) in general involves three steps: feature detection, feature description and classification based the on feature descriptors. Popular feature detection methods ("interest point detectors") are Harris3D [12], Cuboids [6] and Hessian [25]. For descriptors, popular methods are Cuboids [6], HOG/HOF [13], HOG3D [10] and Extended SURF [25]. Most feature descriptors use a bag-of-words visual model that discards all postion information. This reduces the complexity of the features. Recently, [24] combined various feature detection and feature description methods and discovered that there is no universal set of hand-designed features for all datasets. This is a major motivation for moving towards unsupervised feature learning.

[2] have used hand-crafted features for representing video data in their system for the TRECVID 2011 Multimedia Event Detection task. They have extracted key frames from each video and for each key frame, they have extracted local features like SIFT, Color SIFT and Transformed Color Histogram(TCH). Feature vector for the entire video is then computed by averaging over the feature vector of each key frame. Apart from the low-level local features described above, they have also extracted high-level features like PittPatt face detection features, semantic indexing concepts and optical character recognition.

In recent years, several unsupervised learning techniques have been proposed to learn features from the data. Examples of such techniques include models based on restricted Boltzmann machines(RBMs) and deep belief networks. [4] proposed a hierarchical, distributed probabilistic model for unsupervised learning of spatio-temporal features from video data using convolutional restricted Boltzmann machines as a basic processing unit. Their model aggregates over alternating layers of spatial and temporal CRBMs which enables their model to capture long range statistical dependencies in both space and time. [21] proposed a technique based on convolutional gated RBMs that learns latent representations of image sequences from pairs of successive images. Their model captures not only features from static images but motion-sensitive features from pairs of images(consecutive frames in video). The convolutional architecture of their model enables it to extract features from high resolution videos as well.

[14] have used a stacked convolutional neural network along with the ISA to learn features from video data. The features extracted have been used for action recognition task on a wide variety of

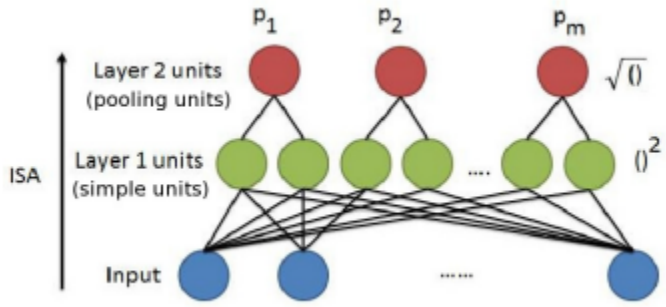

Figure 1: A generic ISA neural network.[14]

datasets like KTH [19], Hollywood2 [16] and UCF [18]. Their system's performance was better than the state-of-the-art performamce on several of these datasets. Our work is an application of their algorithm to the more challenging task of multimedia event detection.

3 Hierarchical Feature Learning

Neural nets have been shown to be efficient in learning to extract features from high dimensional data like video data. In order to extract higher level visual features, we need an invariant model which is able to learn such features hierarchically. We use the approach of Le et al. [14], employing ISA with stacking and convolution for scalability. The following sections build up the model from the bottom-up outlining their approach.

3.1 Standard ISA

The ISA algorithm is a generalization of Independent Component Analysis [23]. ISA automatically extracts features from unlabeled static images by considering a small patch of the image at a time. When implemented using a 3-layer neural network, the first layer consists of the input pattern x^T, the second layer contains square non-linearities and the third layer contains square-root non-linearities. The network learns the weights W connecting the first and second layers, but keeps the weights V, connnecting the second and third layers, fixed in order to account for the second layer neurons' subspace structures.

As such, we obtain the activation for third layer neurons using

$$p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} (\sum_{j=1}^{n} W_{kj} x_j^t)^2} \tag{1}$$

By solving

$$\begin{aligned} &\underset{W}{\text{minimize}} \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^t; W, V) \\ &\text{subject to } WW^T = \mathbf{I} \end{aligned} \tag{2}$$

ISA finds the set of weights W. Here, n is the number of dimensions in the input data, k is the number of neurons in layer 2, m is the number of neurons in layer 3, $\{x^t\}_{t=1}^T$ are whitened examples of input data, $W \in \mathbb{R}^{k \times n}$ are the weights that connect the first layer to the second layer and $V \in \mathbb{R}^{m \times k}$ are the weights that connect the second layer to the third layer. The output of the final layer of the ISA is generally invariant making it quite suitable for extracting features from video data.

3.2 Stacked Convolution ISA

The standard ISA algorithm is not scalable when large inputs such as entire images or videos are fed into the network as input. The non-scalability is due to the necessity of orthogonalization after

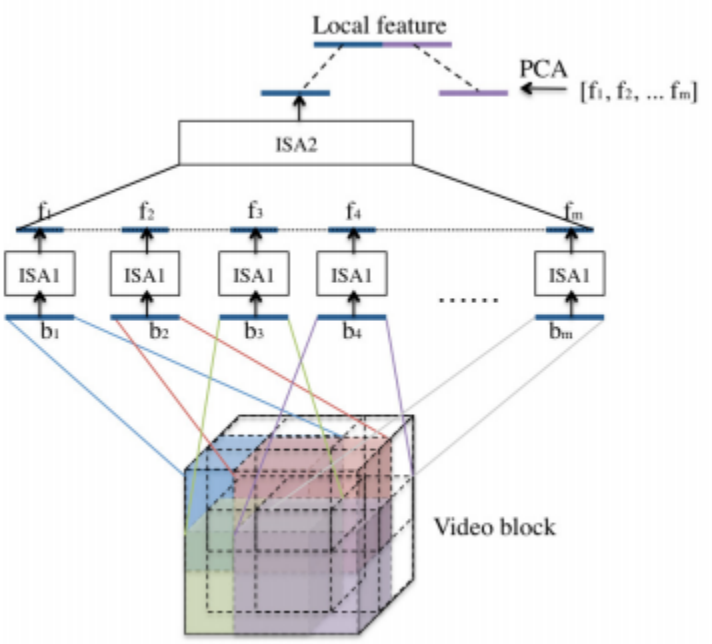

Figure 2: Using stacked convolutional ISA to extract high-level features from video data.[14]

every step in projected gradient descent (see Section 3.4). Since the costs of orthogonalization are proportional to the cube of the dimensions in the input data, using standard ISA on video data (which is high-dimensional) would take a lot of time.

To avoid this problem, ISA is first trained on small patches (should be overlapping patches) of the image. The smaller input sizes make the ISA much more scalable to train as the cost of finding the inverse of the square root of the matrix (Section 3.4) is substantially reduced. The trained ISA is subsequently convolved with the entire image (convolution) and the output is combined to generate a set of features. These first layer features are then fed to PCA as inputs in order to be whitened and reduced in the number of dimensions. The whitened, reduced-dimensional data is then fed as input to a second layer of ISA (stacking). Stacking helps in learning higher level features which can also be done by directly increasing the number of hidden layers in the fundamental neural network structure (ISA in this case), but training such a network will take much longer. Thus, stacking and convolution aid in large-scale deep learning.

3.3 Learning spatio-temporal features

[14] extend the apporach described in Section 3.2 to the video domain. They take a sequence of image patches and flatten them into a vector which is then fed as input to the entire architecture. The sequence of image patches can be interpreted as three-dimensional (the third dimension being the temporal dimension) blocks of data from a video.

Figure 2 shows the application of the architecture to the video domain. [14] have shown that using both higher and lower level features is more useful for action recognition, rather than using either of them alone. Previous literature in event detection also proves the same. Therefore, we use the same approach by combining the intermediate features learned from the first layer with the second layer features and composing them into a single vector.

3.4 Convex Optimization

To learn the weights W for the neural network, [14] utilize batch-projected gradient descent. The objective function, as described in Equation 2, is convex when V is fixed. To keep satisfying the constraint in Eq 2, during each iteration of gradient descent, we are required to update W by pro-

Event	Average Precision SVM	Average Precision RF
Attempting a Board Trick	0.321	0.284
Feeding an Animal	0.244	0.397
Landing a Fish	0.244	0.255
Wedding Ceremony	0.215	0.301
Working on a Woodworking	0.206	0.215

Table 1: Labelwise Average Precision for SVM and Random Forest

jecting it into the constraint set. Minimization of the Frobenius distance based projection objective updates W to $(WW^T)^{(-1/2)}W$.

Solving $(WW^T)^{-1/2}$ requires solving an eigenvector problem, which increases cubically in computational cost. This prevents standard ISA from scaling to high-dimensional data. Fortunately, the problem is mitigated using convolution to reduce the number of dimensions encountered at any one time. The efficiency of PCA is not quite as important, since it is only used twice in total.

4 Classification

The learned spatio-temporal features from the stacked two layer convolutional network are subsequently fed as input to SVM with a chi-square kernel. Since, each video is only one label (event) we do not use multi-label classification. The kernel function is shown below.

$$k(x, y) = \exp\left(-\sum_i \left(\frac{(x_i - y_i)^2}{x_i + y_i}\right)\right) \tag{3}$$

We have also used random forest for classification and compared the results with the results of SVM. Random forest has also been used to get the relative importance of each feature (SVM cannot give such analysis) which has been used for feature analysis.

Both SVM and Random Forest classifiers have been trained for each event class in a one-vs-all classification framework.

5 Evaluation

Receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate) at various threshold settings. ROC curves are a standard technique for evaluating one-vs-all classification tasks. We plot precision vs. recall curves for each class, and the Area Under the Curve (AUC) is the product of the precision and the recall. Typically, we would want to maximize the AUC rather than just precision or recall.

6 Results

The precision recall curves for 5 classes of events using chi-square SVM and Random Forest with 500 trees are shown in 3 and 4. The average precision and recall obtained by both classifiers are shown in the Table 1. Event1 (Attempting a board trick) had the highest average precision with SVM as classifier and Event2 (Feeding an Animal) had the highest average precision with Random Forest as classifier. The mean average precision using SVM with chi-square kernel is 0.25 and the corresponding value for Random Forest is 0.29.

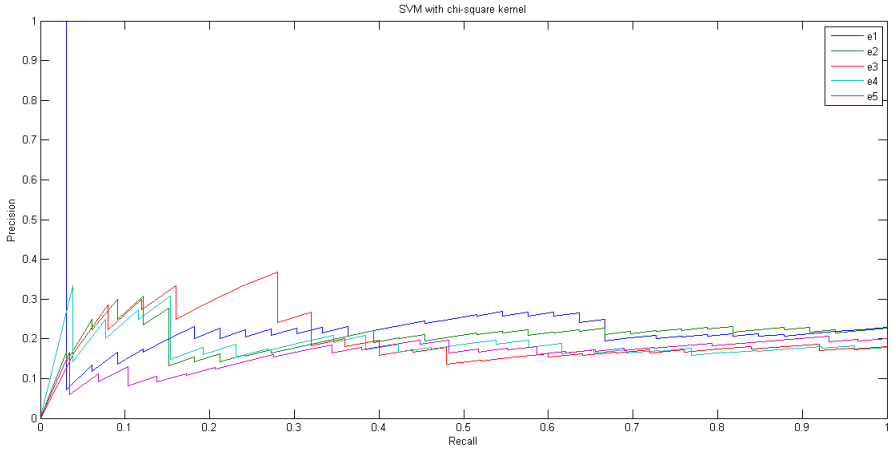

Figure 3: Precision Recall curve for 5 events using SVM

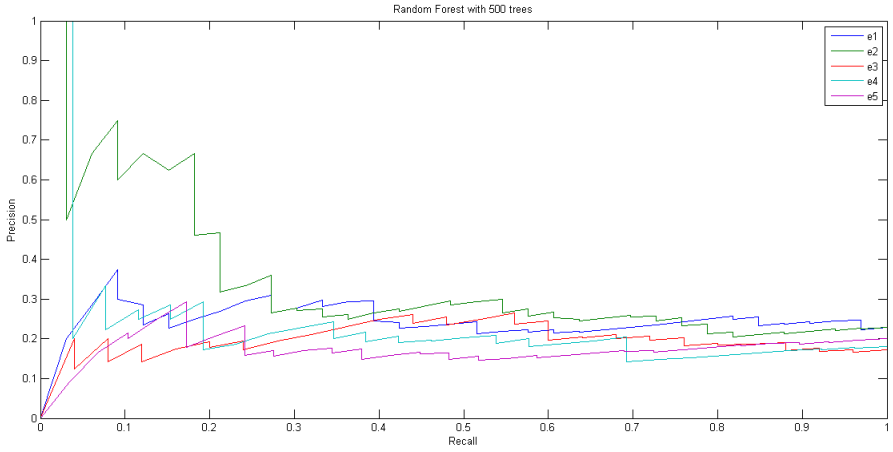

Figure 4: Precision Recall curve for 5 events using Random Forest

6.1 Feature Analysis

It is difficult to do feature analysis with features extracted in an unsupervised manner. However, random forest classifier gives us the relative importance of each feature for each model. Figures 5 and 6 show the relative importance of each feature for random forest models trained for Event 2 (best performance) and Event 5 (worst performance).

The distribution of relative importance of features is more varied in Figure 5 than in Figure 6. Figure 8 shows the number of features with importance value above the mean relative importance for each class. An observable trend in this plot is that well-performing classes have a higher number of features with relative importance above the mean relative importance for that class. Event 2 had the most number of such features whereas event 5 had the least number of such features.

6.2 Fusion of Audio and Video features

This section is the same as that of the audio team.

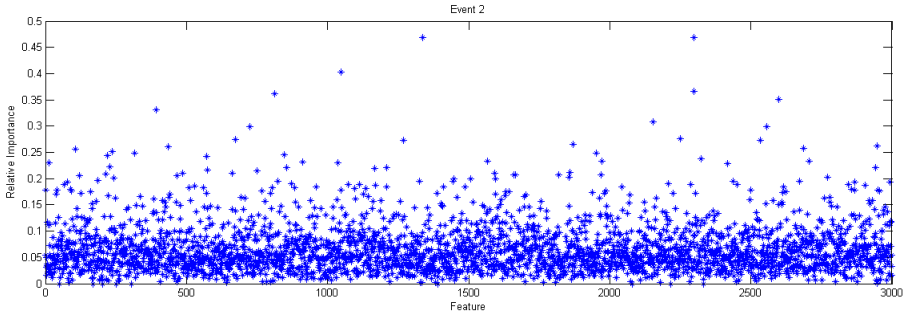

Figure 5: Scatter plot of feature vs relative importance for Event 2

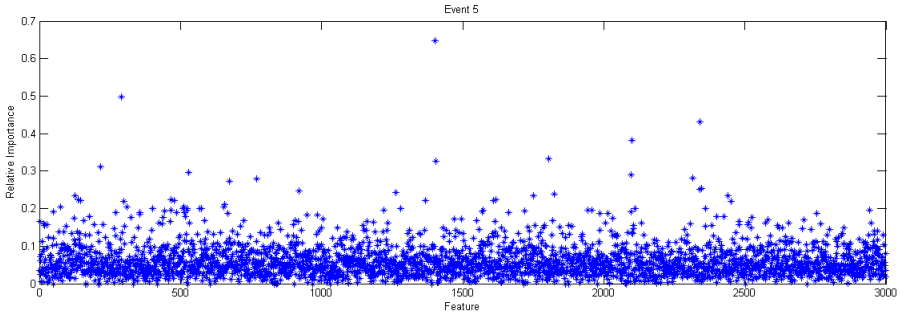

Figure 6: Scatter plot of feature vs relative importance for Event 5

Multimedia is composed of audio and video. For efficient multimedia event detection, we need to use concepts from both audio and video. For example, an event like a 'birthday party' is composed of video concepts like 'birthday cake' and 'faces' and audio concepts like 'cheering'.

The problem of fused classification involves detecting the presence of an event taking into account both audio as well as video features in an input multimedia file. We use the probability returned by SVM and Random Forest for both the audio and video dataset as input features to solve the fused classification problem. (Thus, the number of input features are 4). We then used Logistic Regression for classification. We have tried all combinations of features and report our result on the best feature set (Random Forest audio and video features). Figure 7 shows the precision vs. recall curve for the fusion experiment using logistic regression. It is important to note that the training and test set that we used for the fusion experiment is small.

7 Conclusion

In this report, we applied a method of unsupervised learning to automatically extract features from spatio-temporal data. Using [14] approach, we applied the ISA algorithm combined with deep

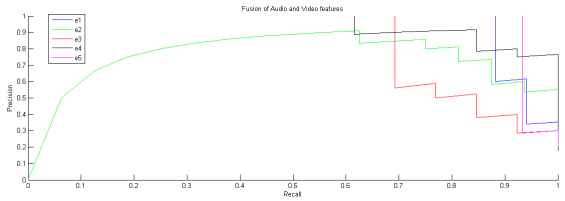

Figure 7: Precision Recall curve using logistic regression after fusion

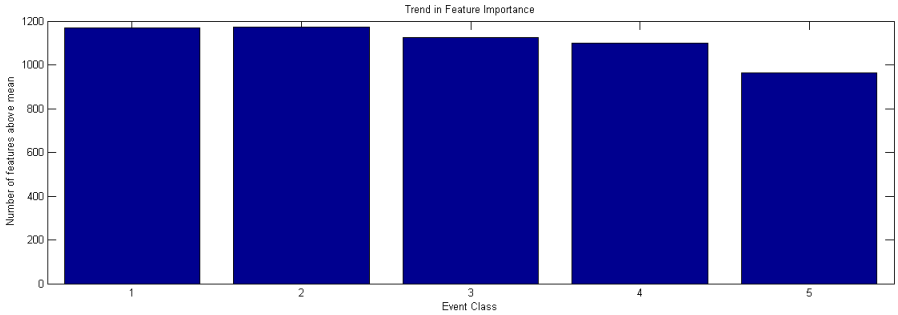

Figure 8: Number of features with importance value greater than mean importance for each class

learning techniques like convolution and stacking to extract features from video data . We then used these features to train SVM and Random Forest classifiers for the purpose of event classification.

All our experiments were performed on a subset of the dataset released for the 2011 TRECVID MED task. Using unsupervised feature learning and SVM with chi-square kernel for classification, we were able to obtain a mean average precision of 0.25. We have also performed classification using Random Forest to obtain a mean average precision of 0.29.

Our results show that unsupervised feature learning can be effective in extracting features from video for event classification. Combining ISA with convolution and stacking helped in scaling the algorithm to high-dimensional video data and extracting hierarchical features from the data. Even though our training data is a subset of the training data released for TRECVID 2011 MED task, we were able to obtain a mean average precision comparable to that of [28]. We have also presented preliminary results of fusion of audio and video features for event classification.

Event Id	Event Name
E001	Attempting a board trick
E002	Feeding an animal
E003	Catching a fish
E004	Wedding ceremony
E005	Working on a wood project
E006	Birthday party
E007	Changing a vehicle tyre
E008	Flashmob gathering
E009	Getting a vehicle unstuck
E010	Grooming an animal
E011	Making a sandwich
E012	Parade
E013	Parkour
E014	Repairing an appliance
E015	Working on a sewing project

Table 2: List of events in the dataset.

Appendix

The 15 events used in our dataset are listed in Table 2. The team working on the audio features comprised of Anurag Kumar and Rohan Ramanath.

References

[1] Multimedia event detection. `http://www.nist.gov/itl/iad/mig/med11.cfm`.

[2] Lei Bao, Shoou-I Yu, Zhen-zhong Lan, Arnold Overwijk, Qin Jin, Brian Langner, Michael Garbus, Susanne Burger, Florian Metze, and Alexander Hauptmann. Informedia@ trecvid 2011. *TRECVID2011, NIST*, 2011.

[3] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153, 2007.

[4] Bo Chen. Deep learning of invariant spatio-temporal features from video. 2010.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[7] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[8] Aapo Hyvèarinen, Jarmo Hurri, and Patrick O Hoyer. *Natural Image Statistics*, volume 39. Springer, 2009.

[9] Lu Jiang, Alexander G Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 449–458. ACM, 2012.

[10] Alexander Klaser and Marcin Marszalek. A spatio-temporal descriptor based on 3d-gradients. 2008.

[11] I. Laptev and T. Lindeberg. Space-time interest points. ICCV, 2003.

[12] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003.

[13] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011.

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[16] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.

[17] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[18] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. ICCV, 2008.

[19] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[20] Cees GM Snoek, Marcel Worring, Jan C Van Gemert, Jan-Mark Geusebroek, and Arnold WM Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430. ACM, 2006.

[21] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *Computer Vision–ECCV 2010*, pages 140–153. Springer, 2010.

[22] Vincent S Tseng, Ja-Hwung Su, Jhih-Hong Huang, and Chih-Jen Chen. Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. *Multimedia, IEEE Transactions on*, 10(2):260–267, 2008.

[23] J Hans van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

[24] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, Cordelia Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, 2009.

[25] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008*, pages 650–663. Springer, 2008.

[26] Shu-Fai Wong and Roberto Cipolla. Extracting spatiotemporal interest points using global information. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[27] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan. Live sports event detection based on broadcast video and web-casting text. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 221–230. ACM, 2006.

[28] Guangnan Ye, I Jhuo, Dong Liu, Yu-Gang Jiang, DT Lee, Shih-Fu Chang, et al. Joint audio-visual bi-modal codewords for video event detection. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 39. ACM, 2012.