Multi-Task Transfer Methods to Improve One-Shot Learning for Multimedia Event Detection

Wang Yan
wyan@sfu.ca
Jordan Yap
jjyap@sfu.ca
Greg Mori
mori@cs.sfu.ca

School of Computing Science
Simon Fraser University
Burnaby, BC, CANADA

Abstract

Learning a model for complex video event detection from only one positive sample is a challenging and important problem in practice, yet seldom has been addressed. This paper proposes a new one-shot learning method based on multi-task learning to address this problem. Information from external relevant events is utilized to overcome the paucity of positive samples for the given event. Relevant events are identified implicitly and are emphasized more in the training. Moreover, a new dataset focusing on personal video search is collected. Experiments on both TRECVid Multimedia Event Detection video set and the new dataset verify the efficacy of the proposed methods.

1 Introduction

The amount of Internet video uploaded by individuals is growing rapidly due to the success of video sharing sites such as YouTube, and leads to numerous challenges in video management and understanding. One of the most important among them is to detect/retrieve a specific event described by a set of exemplar videos. There is a significant amount of work (e.g. [14, 19, 20, 22, 30, 31, 33]) on this topic, and there are also public competitions such as TRECVid Multimedia Event Detection (MED) [23].

Most existing works focus on the setting where quite a few positive examples are available for training an event model. However, in many cases a "query-by-example" setting exists – only one example is provided as the event exemplar. There exist few approaches [17, 18] exploring this problem. TRECVid MED does have evaluation for 0 example (0Ex) learning. However, this is based on a textual description of the event, which is different from the exemplar-based situation examined in this paper.

This paper proposes a one-shot learning method for complex Internet video event detection. In the problem definition, one positive and plenty of negative samples of one event are given as training data, and the goal is to return the videos of the same event from a large video dataset. In addition, we assume samples of other events are available. In this paper, "detection" means to detect videos corresponding to the event of interest from a (large) video dataset, not to localize the event spatially or temporally in a video.

Pages 37.1-37.13
DOI: https://dx.doi.org/10.5244/C.29.37

Fig. 1(a) shows an overview of the proposed methods. The widths of the lines between the one-exemplar event and others represent the inter-event relevance, which is unknown a priori in our problem settings. However, the proposed method can implicity infer the relevance and utilize the most relevant event(s) more in multi-task learning [2], where the shared information from the relevant events helps to build a better model from the one exemplar. The proposed method does not assume the relevance between other events, as indicated by the red line. Although the learning algorithm outputs models of all input events, only that of the one-exemplar event is applied to detect videos of the event of interest from the video set. Fig. 1(b) illustrates other baseline methods evaluated in the experiment using similar graphs. Please refer to Sec. 4 for more details.

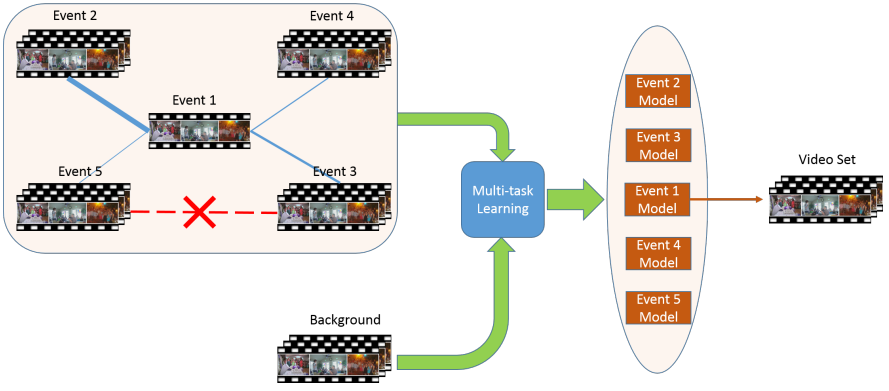
The main contributions of this paper are as follows. First, a new multi-task learning algorithm with implicit inter-task relevance estimation is proposed. Second, the proposed algorithm is applied to address the one-shot learning problem in complex event detection. Last, a new dataset focusing on personal video collection is collected.

The rest of the paper is organized as follows. Sec. 2 discusses the previous work related to the proposed method, which is presented in Sec. 3 in detail. Sec. 4 presents the experimental results and Sec. 5 concludes the paper.

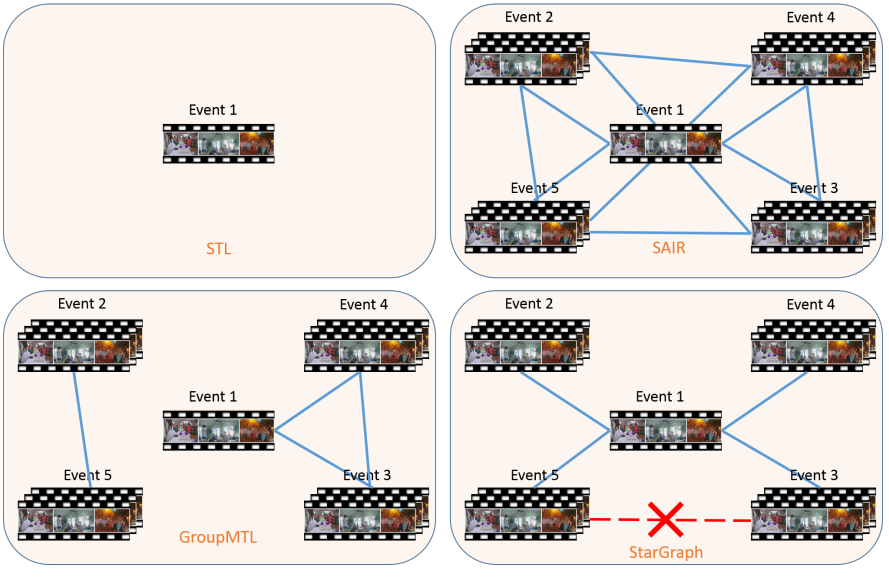
2 Related Work

A comprehensive review of multi-task learning is beyond the scope of this paper, and we will only discuss the relevant applications on action and event modeling. Mahasseni and Todorovic [16] uses squared trace-norm regularized multi-task learning in the multi-view action recognition, where different views are regarded as tasks. Moreover, it identifies the task grouping automatically to encode the non-uniform correlation between tasks. Multi-view action videos taken in controlled environments, e.g. IXMAS [32] and i3DPost [2] are used in the experiments. In contrast, the proposed method does not assume or rely on the clustered structure of the tasks, but utilizes the inter-task relevance directly. Moreover, the tasks in the proposed method correspond to the video events, which are more complex than the action videos, and we develop methods for leveraging other event categories automatically, rather than relying on data of the same event. Zhou et al. [36] represent the task model as the product of a shared matrix and a task-specified vector, thus explicitly models the subspace sharing between tasks. The representation makes the optimization non-convex and computationally intensive. The proposed method does not model the subspace sharing, but considers the similarities between task (event) models directly. Therefore there is not common subspace or feature selection shared by all tasks generally. The series of works by Ma et al. [13, 14, 15] work in the same domain as our proposed method – they also use multi-task learning to train event models on the TRECVid MED dataset. Ma et al. [13] uses the $l_{2,p}$-norm regularizer to select feature dimensions shared by all tasks, while Ma et al. [14] explicitly models the subspace sharing between tasks in a manner similar to Zhou et al. [36], but with a different regularizer. The information sharing in Ma et al. [15] is formulated in a directed way, i.e. the trained video attributes are used to model the event of interest, but the reverse way is not explicitly encouraged in the formulation. The proposed method is different from these three works in several ways. First, the difference to [36] still holds. Second, [13, 14, 15] always use all external data, and sometimes over-estimate the relevance between tasks, while the proposed method can select only a few relevant events automatically. Lastly, [13, 14, 15] have not explored the one-shot learning problem.

Many multi-task methods jointly train classifiers for all tasks, which are assumed to be



(a) Overview of the proposed method – one-shot learning leveraging other categories. Given one example for an event of interest (Event 1), we implicitly infer the relevance between it and other events, and emphasise more on the most relevant ones in the multi-task learning. The learned classifier for the event of interest is applied to retrieve instances from a video test set.



(b) Illustration of other existing methods. Single task learning just uses the event of interest. SAIR allows information sharing between all events. GroupMTL identifies the event group structure and only allows information sharing within groups. StarGraph only allows equally weighted information sharing between the event of interest and other events.

Figure 1: Overview of the proposed method and comparison with existing methods.

related to each other. However, the performance might drop if unrelated tasks are learned together and are forced to share the information. This is a common problem in many applications, and several task grouping methods have been proposed to solve it. Jacob et al. [8] and Zhou et al. [35] consider the distances between task models as the relatedness measure, and an additional term based on within-cluster distance is added in the objective. Kang et al. [9] and Mahasseni and Todorovic [16] split the regularization term in the objective into several, each of which corresponds to the task models within one cluster. In the above methods, the task models are grouped into clusters. In contrast, our proposed method uses the pair-wise task model similarity directly rather than inferring task clusters, which may not exist. Moreover, the goal of existing methods is to maximize within-cluster relatedness for all clusters, while the proposed method only select the tasks most related to the task of interest. Kim [10] and Chen [3] also use the pair-wise model similarity directly in the learning problem, and are more relevant to ours. However, our method implicitly select the most relevant task(s), while their works use all.

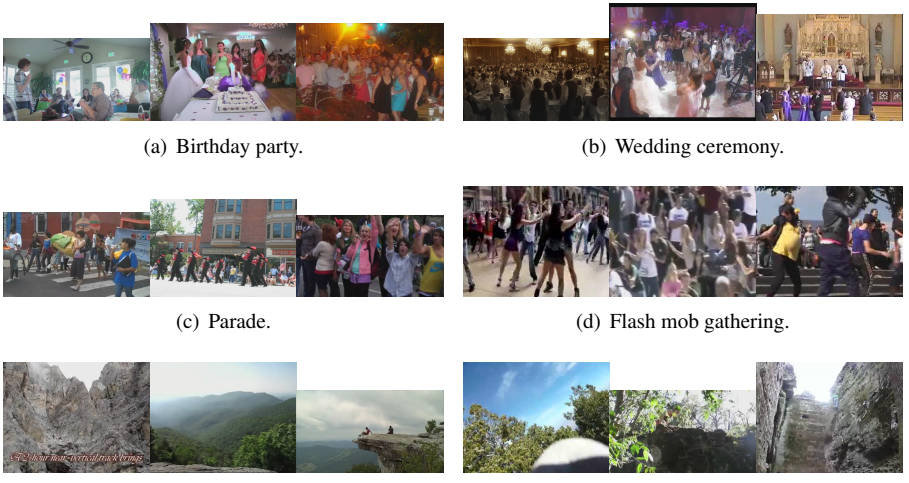
Although it has been widely studied in other applications [6, 26], there are very few works on one-shot learning for video event detection [17, 18]. A query-by-example approach with pre-defined similarity measure is employed by Mazloom et al. [17]. In subsequent work [18], event detection is based on textual features. In our work we learn a classifier, and focus on visual features without additional side information.

There are many video datasets available in the community, e.g. UCF101 [29], HMDB [11], Olympic Sports [21], TRECVid MED [23] and Kodak's Consumer Video Benchmark [12]. However, the following differences make our Vine dataset unique of its kind. First, it focuses on personal video search, and the videos are organized according the subject's identity. TRECVid MED, UCF101, HMDB or Olympic Sports do not have this information. Second, clear video categories are hard to define in the proposed dataset, and the relevance judgement is labeled per video pairs in order to describes the video relevance precisely. In contrast, TRECVid MED, UCF101, HMDB or Olympic Sports collect videos according to pre-defined categories, and the resulting video sets may be biased. While Kodak's Consumer Video Benchmark provides unfiltered consumer videos, it is hard to infer the relevance between videos based on its labeling, which is again limited to pre-defined categories.

3 Proposed Method

We develop an algorithm for learning to retrieve videos of a category given a single positive example video. This one-shot learning setting is challenging, since little information is provided to capture the possible variation in this category. In order to address this issue, we leverage additional data in the form of other video categories.

Most one-shot learning methods use external information, which may help to represent the only sample in a more semantically meaningful way and/or provide some prior on the model. This also holds in the video event detection case, where it is possible to find the relevant events that share the key attributes with the event for which we want to build a model. The video frames from different events in Fig. 2 illustrate this possibility. For example, "wedding ceremony" and "birthday party" both have lots of people sitting or standing still, while "parade" and "flash mob gathering" both have moving crowds, and "hiking" and "rock climbing" both have outdoor scenes. Those shared attributes from other events can potentially help to build a better model of a specific event.



(a) Birthday party. (b) Wedding ceremony.

(c) Parade. (d) Flash mob gathering.

(e) Hiking. (f) Rock climbing.

Figure 2: Video event frame samples.

3.1 Multi-task Learning

Multi-task learning is a broad, active area of research with many different algorithms and approaches. Our method builds on the approach of graph-guided multi-task learning [5]. In this section we describe this approach.

The training set $\{(\mathbf{x}_{ti}, y_{ti}) \in \mathbb{R}^D \times \{-1,+1\}, t = 1,2,\ldots,T, i = 1,2,\ldots,N_t\}$ is grouped into T related tasks, which are further organized as a graph $G =< V,E >$. The tasks correspond to the elements in the vertex set V, and the pairwise relevance between Task t and k are represented by the weight r_{tk} on edges $e_{tk} \in E$. The more relevant the two tasks are, the larger the edge weight is. The graph guided multi-task learning algorithm learns the corresponding T models jointly, by solving the optimization problem

$$\min_{\mathbf{W},\mathbf{b}} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \text{Loss}(\mathbf{w}_t^T \mathbf{x}_{ti} + b_t, y_{ti}) + \lambda_1 \|\mathbf{W}\|_F^2 + \lambda_2 \Omega(\mathbf{W}), \tag{1}$$

where $\mathbf{w}_t \in \mathbb{R}^D$ and $b_t \in \mathbb{R}$ are the model weight vector and bias term of Task t, respectively, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T)$ is the matrix whose columns are model weight vectors, $\mathbf{b} = (b_1, b_2, \ldots, b_t)$, $\|\mathbf{W}\|_F^2 = \text{Trace}(\mathbf{W}^T\mathbf{W})$ is the squared Frobenius norm,

$$\Omega(\mathbf{W}) = \sum_{e_{tk} \in E, t<k} r_{tk} \|\mathbf{w}_t - \mathbf{w}_k\|_2^2 \tag{2}$$

is the graph-guided penalty term. For the significantly relevant tasks, their model weight vectors are forced to be similar due to the large edge weights, and the information could be transferred between relevant tasks. $\text{Loss}(\cdot,\cdot)$ is the loss function. In our work we use logistic loss

$$\text{Loss}(s,y) = \log(1 + \exp(-ys)), \tag{3}$$

which is smooth and leads to an easier optimization problem compared to the hinge loss.

3.2 Multi-task Learning for Event Detection

In the one-shot learning setting for event detection, only one positive sample is available for the specific event. However, a few positive samples are available for other events, and there are also plenty of negative samples for every event. It is hard to learn a good model for the specific event of interest from the only one positive sample. However, due to the potential relevance between events, one can expect a better model by applying multi-task learning to the event with one positive sample and some other events. In the multi-task setting, each of these other events corresponds to one task.

Different external events may share different common "part" with the event of interest. Consider that "birthday party" as the event of interest, and it is relevant to "parade" since both have lots of people inside, and it may be also relevant to "preparing food" due to the food itself. However, there is little relevance between "parade" and "preparing food". Therefore, cluster-based multi-task learning may not be used to learning the event of interest, because not all external events relevant to the event of interest are relevant enough to each other to fit into one cluster. In contrast, graph-guided multi-task does not assume the clustered structure of the tasks, and it a better choice for this task.

Without losing the generality, we assume Event 1 is the event of interest and others are external event in the following. Given the graph in Fig. 1(a), the formulation in (1) is not directly applicable because the pairwise relevance is unknown. However, the min operation can be added to select the most relevant tasks automatically, and they are all equally weighted, i.e. using

$$\Omega(\mathbf{W}) = \sum_{t \in \mathcal{T}_K} \|\mathbf{w}_1 - \mathbf{w}_t\|_2^2 \tag{4}$$

instead of (2), where $\mathcal{T}_K$ is the set of indices corresponding to the minimum K elements in $\{\|\mathbf{w}_1 - \mathbf{w}_t\|_2^2\}_{t=2}^T$. The most relevant tasks and task models are jointly optimised in the training process. The minimum operation can be further replaced by softmin function to make the objective smooth, i.e.

$$\Omega(\mathbf{W}) = -\log \sum_{t=2}^{T} \exp(-\|\mathbf{w}_1 - \mathbf{w}_t\|_2^2). \tag{5}$$

This penalty focuses more on the smallest inter-model distance, which is slightly different from the former one with equally weighted K smallest distances. The experiments show that we can get good results with the smooth penalty. In addition to squared l_2 distance, one can also use the penalty term represented by correlations. The term still focuses more on most relevant tasks, but softmax is used in the representation, i.e.

$$\Omega(\mathbf{W}) = -\log \sum_{t=2}^{T} \exp(\mathbf{w}_1^T \mathbf{w}_t). \tag{6}$$

We use the first option in the following experiments. All penalties in this subsection make the objective non-convex, but one can still get good results empirically. The objective is optimized by Quasi-Newton Soft-Threshold (QNST) method [27].

4 Experimental Results

We evaluate the proposed method by performing one-shot video search experiments on two different video sets, and the mean Average Precision (mAP) is used as the performance measure. The following methods are compared in the experiments.

STL stands for Single Task Learning, which optimizes the logistic loss (3) plus the squared 2-norm $\lambda\|\cdot\|^2$ regularizer. It only uses the one exemplar and background to learn the event model, and performs only a single task at one time. It serves as a "one example only" baseline method, with no side information. The only parameter λ is tuned within $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.

SAIR [14] is a multi-task learning method. In one-shot learning of one event model, all other events are considered as the "concepts", and the corresponding exemplar videos are used accordingly. The parameter p is empirically set to 1 and α is tuned within $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The authors' code is used.

GroupMTL [9] is a multi-task learning method which learns task groups, and the information is shared between tasks within the same group only. The number of groups g is set to 10, and the regularizer coefficient λ is tuned within $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The authors' code is used.

StarGraph stands for the graph-guided multi-task learning with equally weighted penalties, i.e. problem (1) with penalty term

$$\Omega(\mathbf{W}) = \sum_{t=2}^{T} \|\mathbf{w}_1 - \mathbf{w}_t\|_2^2. \tag{7}$$

The parameters λ_1 and λ_2 are tuned within $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and $\{10^{-3}, 10^{-1}, \ldots, 10^{3}\}$, respectively. It is an (over)-simplified version of the proposed StarGraphSoftMin.

StarGraphSoftMin is the proposed method, and stands for the graph-guided multi-task learning with softmin, i.e. problem (1) with penalty term (5). The parameters λ_1 and λ_2 are tuned within the same ranges as StarGraph.

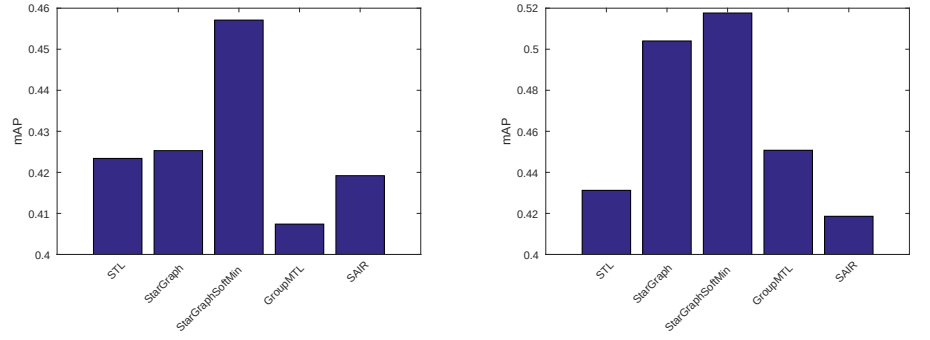
Fig. 1(a) illustrateds the proposed methods, which allows weighted information sharing between the event of interest and other events. As a comparison, Fig. 1(b) shows information sharing between tasks using similar symbols. STL just uses the event of interest, and there is no sharing. SAIR allows information sharing among all events, while GroupMTL identifies the group structure within events and only allows information sharing within groups. StarGraph only allows equally weighted information sharing between the event of interest and other events.

4.1 TRECVid MED + UCF101

We first test all methods on the 20 TRECVid MED13 pre-specified events[1]. The Event Kits set is used as the training set, which has 100 positive video exemplars and 2-10 near-miss videos for each event, and 4992 background videos (Event Kits BG) considered as negative samples. The testing set consists of all event videos (16-234/event) and 500 randomly sampled background videos from the MEDTest set. For the multi-task methods, UCF101 [29] actions are used as the external data. It has 13320 video clips for 101 action categories, and at least 100 clips/action.

The experiments are performed in the one-shot learning setting, which has been described in Sec. 1. For one MED event, it is presumed that only one exemplar video is available, one should build an event model based on the exemplar, Event Kits BG and UCF101 videos of all 101 actions. For multi-task methods, the external task is defined as learning

[1] Birthday party, Changing a vehicle tire, Flash mob gathering, Getting a vehicle unstuck, Grooming an animal, Making a sandwich, Parade, Parkour, Repairing an appliance, Working on a sewing project, Bike trick, Cleaning an appliance, Dog show, Giving directions, Marriage proposal, Renovating a home, Rock climbing, Town hall meeting, Winning race without a vehicle, Working on a metal crafts project.



(a) TRECVid MED. (b) Vine.

Figure 3: mAPs on two datasets.

the model of one action using UCF101 videos and Event Kits BG. These tasks are jointly learned with the primary task, i.e. learning the MED event model using the exemplar video and Event Kits BG, and the information sharing helps the primary task training.

In order to make the maximum use of the rare positive videos, the testing set is augmented by the rest 99 training videos of the event of interest. The learned model is used to score and rank the testing set, and AP is computed given that all videos of other events as well as the background are considered as negative. This setting will lead to 100 AP values for one event if every positive video is used as the only exemplar in turn, and the mean AP (mAP) is reported in the following experiments. The one-shot learning setting makes it impossible to select parameters by cross-validation, and therefore the parameters are tuned per query for all methods.

There might be a concern on combining the UCF101 and TRECVid Event Kits BG videos as the external task, which may lead to meaningless or misleading models. However, it just changes the way of training the model, the testing is still performed on pure TRECVid dataset. If the the external task models were bad, they would have probably dragged down the performance at testing due to the information sharing. However, we do not observe such a phenomenon in the experiments.

The low-level feature used is based on the pre-trained deep models [28] from still images. Similar to [34], deep feature is extracted from the central patch of sampled frames, and is then aggregated into video-level feature by temporal pooling and l_2 normalization. Last, PCA is applied to the video-level feature.

Fig. 3(a) shows the mAPs of all methods. StarGraphSoftMin output performs all other methods. It outperforms StarGraph since the latter treat all external task equally, and is prone to be affected by irrelevant tasks. However, StarGraph is still a little bit better than STL thanks to the multi-task learning. SAIR does not get very good performance, and its $l_{2,p}$-norm-based loss may not be suitable for a classification problem. It is likely that GroupMTL cannot identify relevant events reliably given only one exemplar, and fails in this experiment. As a reference, the random mAP on this dataset is 0.083, which is far below the numbers reported in Fig. 3(a). In order to check the consistency of the improvement across different events, we compare the mAP of different methods per event. The numbers of events which have greater, equal or less mAPs using StarGraphSoftMin, compared to other methods, are reported in Table 1. The proposed method outperforms other methods on most of the events.

	STL	StarGraph	GroupMTL	SAIR
Greater	20	20	20	15
Equal	0	0	0	0
Less	0	0	0	5

Table 1: Numbers of TRECVid events which have greater, equal or less mAPs using StarGraphSoftMin, compared to other methods.

(a) (b) (c) (d) (e)

Figure 4: Vine video samples. (b) is relevant to query (a), while other videos are irrelevant.

4.2 Vine + UCF101

Vine Dataset The TRECVid dataset is diverse and challenging, but it focuses on the detection of pre-defined events, not general video search. Moreover, it mixes videos from different authors together, and cannot be used as a benchmark for personal video seach. We collect a new personal video dataset from video sharing website Vine.co to address these two issues. We search for users with randomly generated first or last names [1], and randomly choose one from the search result. For each user, only the originally posted videos are collected, and re-posted videos from other users are discarded. It makes sure that all videos are from the subject. There are 18 subjects and 1447 videos in total, and each subject has 10-212 videos. The video length is 2-6s. Most of the videos are one of a kind, i.e. there is no other relevant videos from the same subject. Besides them, there are small sets of videos have relevant counter-parts, and the queries are randomly selected from these videos. There are 200 queries, and each subject has 3-39 queries. We further find that it is hard to categorize the queries since the category boundary is vague. For example, Video A and C are both relevant to Video B, but they may not belong to the same category because Video A may be irrelevant to Video C. This happens frequently with videos having multiple "concepts" inside. Therefore, the binary relevance judgement as groundtruth is manually labeled per video pair rather than automatically generated based on category label. Fig. 4 shows some videos from one subject. The first video is the query, and the second video is labeled as relevant, while the rest videos are labeled as irrelevant to the query. We plan to release this dataset publicly.

The experiments are also performed in the one-shot learning setting. Each query is in

	STL	StarGraph	GroupMTL	SAIR
Greater	141	94	129	148
Equal	46	55	45	37
Less	13	51	26	17

Table 2: Numbers of Vine queries which have greater, equal or less APs using StarGraph-SoftMin, compared to other methods.

turn used as the only one exemplar video is available to build the event model. Event Kits BG is again used as the negative training samples, for both the primary and external tasks, which are the same as the ones in TRECVid MED + UCF101 experiment. The testing is restricted within the same subject, and the experiment simulates the video search within the personal collection. The parameters are also tuned per query for each method, and the mAPs are reported.

The low-level feature used is based on Histograms of Oriented Gradients (HOG) [4]. Frames are sampled from a video at 4FPS, and are then resized such that the longer side has 640 pixels. HOG are extracted from dense patches in the frames, with patch stride and cell size of 18 pixels, and each patch consists of 2×2 cells. Then, the HOG features are compressed to 32 dimensions by PCA, and Fisher vectors [24, 25] are extracted from the bag of compressed HOG features from one video, with a 256 component Gaussian Mixture Model. 2-level spatial pyramid with branch factor 2×2 is applied to the frame coordinate, and the Fisher vectors extracted from all regions are concatenated. Last, the concatenated vector is projected to a low-dimensional space by PCA once more.

Fig. 3(b) shows the mAPs of all methods. Similar patterns between StarGraphSoftMin, StarGraph and STL are observed, and again show the effectiveness of the proposed method. As a reference, the random mAP is 0.091. Moreover, the numbers of queries which StarGraphSoftMin has greater, equal or less APs compared to other methods are reported in Table 2. The proposed methods is better or on par with other methods on most of the videos.

5 Conclusion

This paper shows that multi-task learning can improve one-shot learning performance in complex event detection, which is an important but seldom addressed problem. We developed a method for implicitly selecting related tasks, requiring no additional labeling for the one example from the event in question. The method only require a single positive example of the event of interest, and leverage examples from other categories. On both the TRECVid MED dataset and a personal video dataset collect by ourselves, we show empirically that these can lead to an effective approach to building a more accurate event model.

Acknowledgement

This research was supported by NSERC.

References

[1] Behind the name. URL http://www.behindthename.com/random/.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] Xi Chen, Qihang Lin, Seyoung Kim, Jaime Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse learnin. In *UAI*, 2011.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[6] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[7] Nikolaos Gkalelis, Hansung Kim, Adrian Hilton, Nikos Nikolaidis, and Ioannis Pitas. The i3dpost multi-view and 3d human action/interaction database. In *Visual Media Production, Conference for*, 2009.

[8] Laurent Jacob, Jean-Philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *NIPS*, 2009.

[9] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011.

[10] S. Kim, K. A. Sohn, and E. P. Xing. A multi-variate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):204–212, 2009.

[11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

[12] Alexander Loui, Jiebo Luo, Shih-Fu Chang, Dan Ellis, Wei Jiang, Lyndon Kennedy, Keansub Lee, and Akira Yanagawa. Kodak's consumer video benchmark data set: concept definition and annotation. In *Multimedia Information Retrieval, International workshop on*, 2007.

[13] Zhigang Ma, Yi Yang, Yang Cai, Nicu Sebe, and Alexander G Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM MM*, 2012.

[14] Zhigang Ma, Yi Yang, Nicu Sebe, Kai Zheng, and Alexander G Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *TMM*, 15(7): 1628–1637, 2013.

[15] Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, and Alexander G Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.

[16] Behrooz Mahasseni and Sinisa Todorovic. Latent multitask learning for view-invariant action recognition. In *ICCV*, 2013.

[17] Masoud Mazloom, Amirhossein Habibian, and Cees GM Snoek. Querying for video events by semantic signatures from few examples. In *ACM MM*, 2013.

[18] Masoud Mazloom, Xirong Li, and Cees GM Snoek. Few-example video event retrieval using tag propagation. In *Multimedia Retrieval, International Conference on*, 2014.

[19] Gregory K. Myers, Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ramakant Nevatia, Chen Sun, Amirhossein Habibian, Dennis C. Koelma, Koen E. A. van de Sande, Arnold W. M. Smeulders, and Cees G. M. Snoek. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 25:17–32, 2014.

[20] Pradeep Natarajan, Shuang Wu, Shiv N. P. Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[21] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[22] Sangmin Oh, Scott McCloskey, Ilseo Kim, Arash Vahdat, Kevin J Cannons, Hossein Hajimirsadeghi, Greg Mori, AG Amitha Perera, Megha Pandey, and Jason J Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, 25:49–69, 2014.

[23] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2011 — an overview of the goals, tasks, data, evaluation mechansims and metrics. In *TRECVID*, 2011.

[24] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[25] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[26] Tomas Pfister, James Charles, and Andrew Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *ECCV*, 2014.

[27] Mark Schmidt. *Graphical model structure learning with l_1-regularization*. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA, 2010.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[30] Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[31] Wei Tong, Yi Yang, Lu Jiang, Shoou-I Yu, ZhenZhong Lan, Zhigang Ma, Waito Sze, Ehsan Younessian, and Alexander G. Hauptmann. E-lamp: integration of innovative ideas for multimedia event detection. *Machine Vision and Applications*, 25:5–15, 2014.

[32] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *ECCV*, 2010.

[33] Yi Yang, Zhigang Ma, Zhongwen Xu, Shuicheng Yan, and Alexander G Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013.

[34] Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144*, 2015.

[35] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, 2011.

[36] Qiang Zhou, Gang Wang, Kui Jia, and Qi Zhao. Learning to share latent tasks for action recognition. In *ICCV*, 2013.