Improving event detection using related videos and Relevance Degree Support Vector Machines

Christos Tzelepis, Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris
Information Technologies Institute/CERTH
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
{tzelepis, gkalelis, bmezaris, ikom}@iti.gr

ABSTRACT

In this paper, a new method that exploits *related* videos for the problem of event detection is proposed, where *related* videos are videos that are closely but not fully associated with the event of interest. In particular, the Weighted Margin SVM formulation is modified so that related class observations can be effectively incorporated in the optimization problem. The resulting Relevance Degree SVM is especially useful in problems where only a limited number of training observations is provided, e.g., for the EK10Ex subtask of TRECVID MED, where only ten positive and ten related samples are provided for the training of a complex event detector. Experimental results on the TRECVID MED 2011 dataset verify the effectiveness of the proposed method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Theory

Keywords

Video event detection, Very few positive samples, Relevance Degree Support Vector Machines.

1. INTRODUCTION

High-level (or complex) event detection in video has recently received considerable attention in a wide range of applications such as video organization, annotation and retrieval [1]. This interest is motivated from recent studies in psychophysics indicating that high-level events play an essential role in structuring our memories and recalling our experiences [2]. That is, it is expected that an event-based organization of video content can significantly contribute to bridging the semantic gap between human and machine interpretations of multimedia.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502176.

There are numerous challenges associated with the detection of events in videos, related with the inherent complexity of events, the variability of videos of a given event that are captured with different video cameras, and other. To deal with these difficulties, usually a large-scale annotated dataset of training videos is used for capturing as much as possible of the variability in the appearance of each event. However, the creation of such a dataset is a tedious and expensive procedure, or sometimes not feasible, e.g., in cases of rare events. To this end, in this paper we investigate the problem of event detection where only a limited number of positive and related (i.e., videos that are closely related with the event, but do not meet the exact requirements for being a positive event instance [3]) event videos are provided. Specifically, the Weighted Margin SVMs (WMSVMs) presented in [4] are modified using an appropriate slack normalization function, so that event information in the related videos can be effectively exploited in the resulting Relevance Degree SVM (RDSVM) formulation.

The paper is organized as follows. In Section 2 related work is reviewed, and in Section 3 RDSVMs are described. Results of the application of RDSVMs to the TRECVID MED 2011 dataset are provided in Section 4, while conclusions are drawn and future work is discussed in Section 5.

2. RELATED WORK

Only a limited number of works have been reported that deal with the problem of event detection in video using a few training observations [3]. Most of them are in the context of the EK10Ex TRECVID MED subtask, where only 10 positive and 10 related videos are provided for each target event. For instance, in [5, 6] various low-level (e.g., SIFT, MoSIFT, etc), intermediate-level (e.g., semantic model vectors related to the TRECVID SIN task), and textual (OCR, ASR, etc) features are used to learn the events. In [7], an auxiliary set containing videos of several semantic concepts is exploited to leverage additional event information. However, differently from our method, none of these works takes full advantage of the related videos provided in the EK10Ex subtask; instead, related samples of each target event are either excluded from the training procedure, or are treated as purely positive or negative instances.

The exploitation of related videos can be achieved using a knowledge adaptation technique, such as WMSVMs [4], or maximum margin discriminant analysis [8]. Typically, these methods derive a weight for each positive observation in the training set and utilize this weight in their optimization criterion. To the best of our knowledge, the applicability of

such techniques to video event detection problems such as that of the EK10Ex subtask has not been investigated.

3. SUPPORT VECTOR MACHINES WITH RELEVANCE DEGREES

Let $\mathcal{X} = \{(\mathbf{x}_i, y_i)|i = 1, ..., N\}$ be an annotated dataset of N observations, where $\mathbf{x}_i \in \Re^L$ is the feature vector representation of the i-th observation in the L-dimensional space with label $y_i \in \{0, \pm 1\}$ denoting that the i-th observation is a positive ($y_i = +1$), a negative ($y_i = -1$), or a related ($y_i = 0$) instance of the class. To allow the use of the RDSVM, the above problem is reformulated as $\mathcal{X} = \{(\mathbf{x}_i, y_i, u_i)|i = 1, ..., N\}$, where $y_i \in \{\pm 1\}$, and $u_i \in \{0, 1\}$ is the so-called relevance label denoting that the i-th observation is a true ($u_i = 0$) or a related ($u_i = 1$) instance of the class y_i.

For the exploitation of the related observations, the SVM formulation presented in [4] is utilized. In [4], each training sample $\mathbf{x}_i$ is associated with a confidence value v_i, and a monotonically increasing function $g(v_i)$ (called slack normalization function) is used to weight each slack variable ξ_i. In this way, the support vectors (SVs) that are associated with a higher confidence value have greater contribution to the computation of the separating hyperplane. Here, we modify this function so that only related class observations are associated with a confidence value $c_i \in (0, 1]$ (called hereafter relevance degree) indicating the degree of relevance of the i-th observation with the class it is related. That is, g is defined as follows:

$$g(u_i) = \begin{cases} 1 & \text{if } u_i = 0, \\ c_i & \text{if } u_i = 1. \end{cases} \quad (1)$$

We then wish to minimize

$$Q(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} g(u_i)\xi_i, \quad (2)$$

subject to the following constraints

$$y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ i = 1...N, \quad (3)$$

where $\mathbf{w} \in \Re^L$, $\xi_i = \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)$, $b \in \Re$ and $C \in \Re_+$ are the weight vector, the slack variables, the bias and the penalty term, respectively. Moreover, φ denotes a mapping from the input space $\Re^L$ into a new, high-dimensional feature space $\mathcal{F}$ [9].

The corresponding primal Lagrangian is given from

$$L_P(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} g(u_i)\xi_i - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) - 1 + \xi_i] - \sum_{i=1}^{N} \beta_i\xi_i, \quad (4)$$

where $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_N]^T$, $\boldsymbol{\beta} = [\beta_1, ..., \beta_N]^T$ are the corresponding non-negative Lagrange multipliers.

Using the Karush-Kuhn-Tucker (KKT) conditions [9] the dual formulation is retrieved:

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j K_{i,j}, \quad (5)$$

subject to the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad (6)$$

$$0 \leq \alpha_i \leq g(u_i)C,\ i = 1, ..., N, \quad (7)$$

where K_{ij} is a Mercer kernel such that $K_{ij} = \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j)$ [9]. This is a quadratic programming problem which can be solved using an appropriate optimization technique (e.g. see [4]).

An unlabelled test observation is then classified according to $y_t = \text{sgn}(\sum_{i\in S} \alpha_i y_i K_{it} + b)$, where S is the set of indices corresponding to the SVs, i.e., the training observations for which $\alpha_i \neq 0$ [9]. From the above equation we observe that the larger the α_i, the more significant is the contribution of the associated SV in the classification process. Therefore, using the above RDSVM formulation (2) and as shown by (7), the contribution of the related samples in the computation of the separating hyperplane can be regulated using appropriate relevance degrees c_i.

4. APPLICATION TO VIDEO EVENT DETECTION

4.1 Dataset description

The large-scale video dataset of the TRECVID MED 2011 task is used for the application of RDSVM in the problem of event detection. This dataset consists of 13113 development and 32037 test videos belonging to the "rest of the world" class, or to one out of 10 target event classes, namely: "Birthday party" (E006), "Changing a vehicle tire" (E007), "Flash mob gathering" (E008), "Getting a vehicle unstuck" (E009), "Grooming an animal" (E010), "Making a sandwich" (E011), "Parade" ("E012), "Parkour" (E013), "Repairing an appliance" (E014), "Working on a sewing project" (E015). Example keyframes for five of them are shown in Fig. 1.

4.2 Feature vector representation

A semantic model vector approach is used for the representation of each video, similarly to the approach described in [10]. In particular, each video signal is decoded and a 1×3 spatial pyramid decomposition scheme is applied to the keyframes. The dense sampling strategy along with the opponentSIFT interest point descriptor are then applied to derive a set of feature vectors in $\Re^{384}$ for each pyramid cell, as well as the entire frame, and the k-means algorithm is utilized to construct a Bag-of-Words (BoW) model of 1000 visual words for each cell. Concatenating the BoW models, each keyframe is represented with a BoW feature vector in $\Re^{4000}$.

Subsequently, a set of L pre-trained concept detectors, $\mathcal{G} = \{g_k : \Re^{4000} \mapsto [0, 1]|k = 1, ..., L\}$ is used to provide a model vector representation of the p-th keyframe of the i-th video $\mathbf{x}_{i,p} = [x_{i,p,1}, ..., x_{i,p,L}]^T$. That is, $x_{i,p,k}$ is the response of the concept detector g_k concerning the p-th keyframe of the i-th video, expressing the degree of confidence (DoC) that the k-th concept is depicted in the keyframe. In order to derive a model vector representation $\mathbf{x}_i$ of the i-th video, the model vectors at keyframe level are averaged, $\mathbf{x}_i = \sum_{p=1}^{T_i} \mathbf{x}_{i,p}$, where T_i is the length of the i-th video in keyframes.

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

Figure 1: Example keyframes for 5 of the target events: (a) birthday party, (b) changing a vehicle tire, (c) flash mob gathering, (d) getting a vehicle unstuck, (e) grooming an animal.

The above concept detectors have been designed using linear SVMs and the TRECVID SIN 2012 dataset [3], which contains annotated videos for $L = 346$ concepts.

4.3 Experimental setup

The training of each target event detector is performed using 10 positive, 10 related and 70 negative videos, selected from the development set (except for the events E007 and E015, where only 6 and 2 related videos are provided, respectively). That is, a setting similar to the EK10Ex one of the TRECVID MED 2012 task is used [3].

For evaluation purposes, the proposed approach is compared with three state of the art experimental scenarios of the literature:

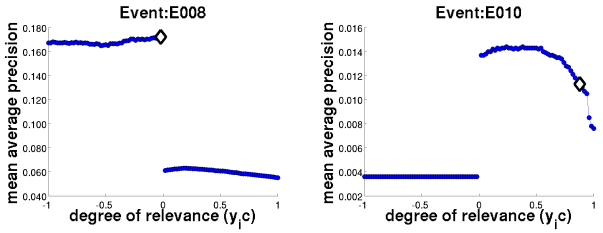
1. **Scenario A (noRelatedSamples)**: Related videos are excluded from this experiment. That is, only 10 positive videos are used for learning each target event, as e.g. in [5].
2. **Scenario B (relatedAsPositive)**: Related videos are used as positive examples (i.e., $y_i = +1$ and $u_i = 0$), as e.g. in [11].
3. **Scenario C (relatedAsNegative)**: Related videos are used as negative examples of each target event (i.e., $y_i = -1$ and $u_i = 0$).

The proposed approach is used in two experimental settings, as follows:

1. **Scenario D (relatedAsWeightedPositive)**: Related videos are used as related positive examples (i.e., $y_i = +1$ and $u_i = 1$) in RDSVM.
2. **Scenario E (relatedAsWeightedNegative)**: Related videos are used as related negative examples (i.e., $y_i = -1$ and $u_i = 1$) in RDSVM.

In all above settings, two different approaches to treating the related samples are tested: i) all 10 related samples are used for training the SVM or RDSVM, ii) the median model vector of the 10 related samples is computed and only the 5 that are closer to that median are used for training.

The proposed RDSVM has been implemented using a modified version of LIBSVM [12]. The radial basis function (RBF) kernel ($K_{ij} = exp(-\| \mathbf{x}_i - \mathbf{x}_j \|^2 / 2\sigma^2)$, where $\sigma \in \Re_+$ is the scale parameter) has been used. Moreover, the relevance degrees for all related videos are set to the same constant value $c \in (0, 1]$, i.e., $c_i = c \; \forall \, i$. The optimization of the parameters (C, σ) for the scenarios A, B and C is performed using a grid search on a 10-fold cross-validation procedure, where at each fold the training set is split to 90% learning set and 10% validation set. For scenarios D and E, the (C, σ) values already selected for scenarios B and C, respectively, are used and the optimum value of c is selected again by cross-validation searching through the interval $(0, 1]$ with a step size equal to 0.02.



(a) E008 (b) E010

Figure 2: Selection of the relevance degrees for 2 target events using cross-validation.

4.4 Results and discussion

Table 1 shows the performance of the proposed method, in terms of average precision (AP), for each experimental scenario and target event. Moreover, for each experimental scenario, the mean average precision (MAP) across all target events is shown in the last column. From these results we observe that the way the related videos should be treated (i.e., as positive or negative samples) depends on the particular event under consideration. It also depends heavily on the actual related samples that are employed. Overall, in terms of MAP, we see that scenarios D and E, which use RDSVM, outperform scenarios A-C, which represent the current state of the art. Moreover, when related samples are pruned using the median clustering, scenarios B-E consistently provide better results. In this case, overall, the best performance is achieved by scenario E.

In Fig. 2, we show the effect of the relevance degree c in the detection performance for two target events (E008 and E010) when related samples are pruned using median clustering as discussed above. In particular, for scenarios D and E we learn the events for different values of c in the range $(0, 1]$ with a step size equal to 0.02, and compute the corresponding MAP achieved in the evaluation set. For the sake of compactness, the results for both scenarios D and E concerning the same target event are depicted in the same diagram. This is achieved using the quantity $y_i c$ (instead of c) in the horizontal axis of the diagrams.

In conclusion, contrary to the most methods in the literature that use the related samples as either purely negative or positive (e.g. see [11]), the proposed method allows for

Table 1: Evaluation of event detection approaches on the TRECVID MED 2011 dataset.

	Target Events										
Experimental Scenarios	*E006*	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	*MAP*
Scenario A	0.0121	0.0076	0.0772	0.0190	0.0035	0.0137	0.0212	0.0615	0.1191	0.0083	0.0343
Full set of 10 randomly selected related samples											
Scenario B	0.0124	0.0317	0.0933	0.0205	0.0027	0.0080	0.0257	0.0551	0.0867	0.0082	0.0344
Scenario C	0.0210	0.0149	0.0613	**0.0299**	**0.0033**	**0.0101**	0.0159	0.0486	**0.1358**	0.0068	0.0348
Scenario D	0.0123	**0.0393**	**0.1197**	0.0205	0.0029	0.0080	**0.0264**	**0.0620**	0.0894	0.0059	**0.0386**
Scenario E	**0.0296**	0.0149	0.0658	**0.0299**	**0.0033**	0.0087	0.0201	0.0507	0.1357	**0.0089**	0.0368
Subset of 5 related samples using the median clustering method											
Scenario B	**0.0466**	0.0060	0.0551	0.0218	0.0076	**0.0098**	**0.0423**	0.0559	0.1249	0.0082	0.0378
Scenario C	0.0117	**0.0119**	0.1672	0.0377	0.0036	0.0085	0.0157	0.0881	0.0631	0.0068	0.0414
Scenario D	0.0447	0.0060	0.0573	0.0218	**0.0113**	**0.0098**	0.0422	0.0594	**0.1357**	0.0059	0.0394
Scenario E	0.0117	0.0118	**0.1721**	**0.0379**	0.0036	0.0092	0.0222	**0.0876**	0.0633	**0.0089**	**0.0428**
Selection D,E	0.0447	0.0118	0.1721	0.0379	0.0113	0.0098	0.0222	0.0876	0.0633	0.0059	0.0467

a more careful treatment of the related samples, which can provide a considerable performance improvement.

Finally, we should note that despite the limited number of positive training videos, for 7 out of 10 target events the best way of treating the related samples (i.e., as negative or positive) in RDSVM could be correctly decided automatically at the RDSVM training stage, by looking at the average of the AP values attained during cross-validation. For instance, in Fig. 2 the signed value of c that was selected in this way is shown using a diamond marker. The overall results of this automatic selection are shown in the last row of Table 1 (named "Selection D, E"), indicating that this process can lead to further improvement of the overall detection results in terms of MAP.

5. CONCLUSIONS AND FUTURE WORK

A method that effectivelly exploits related class observations for the detection of events in videos was proposed in this paper. The applicability of the method was verified using the TRECVID MED 2011 dataset, where similar to the EK10Ex subtask of MED 2012 only a limited number of positive videos were used during the training stage.

In the future, we plan to combine the proposed method with LSSVMs [13] so that relevance degrees can be exploited at the subclass level (i.e., using different values of c_i for different subclasses of the related samples).

6. ACKNOWLEDGMENT

This work was supported by the European Commission under contracts FP7-287911 LinkedTV, FP7-600826 ForgetIT and FP7-318101 MediaMixer.

7. REFERENCES

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, pp. 1–29, 2012.

[2] N. R. Brown, "On the prevalence of event clusters in autobiographical memory," *Social Cognition*, vol. 23, no. 1, 2005.

[3] P. Over, G. Awad, M. Martial, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, and A. F. Smeaton, "TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. of TRECVID 2012*. NIST, USA, 2012.

[4] X. Wu and R. Srihari, "Incorporating prior knowledge with weighted margin support vector machines," in *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*. ACM, 2004, pp. 326–333.

[5] M. Akbacak, R.C. Bolles, J. B. Burns, M. Eliot, et al., "The 2012 SESAME multimedia event detection (MED) system," in *Proc. TRECVID 2012 Workshop*, Nov. 2012.

[6] H. Cheng, J. Liu, S. Ali, O. Javed, et al., "Sri-sarnoff AURORA system at TRECVID 2012 multimedia event detection and recounting," in *Proc. TRECVID 2012 Workshop*, Nov. 2012.

[7] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G Hauptmann, "Knowledge adaptation for Ad Hoc multimedia event detection with few exemplars," in *Proc. 20th ACM Int. Conf. on Multimedia*. ACM, 2012, pp. 469–478.

[8] W. Zheng, C. Zou, and L. Zhao, "Weighted maximum margin discriminant analysis with kernels," *Neurocomputing*, vol. 67, pp. 357–362, 2005.

[9] V. Vapnik, *Statistical learning theory*, New York: Willey, 1998.

[10] A. Moumtzidou, N. Gkalelis, P. Sidiropoulos, M. Dimopoulos, S. Nikolopoulos, S. Vrochidis, V. Mezaris, and I. Kompatsiaris, "ITI-CERTH participation to TRECVID 2012," in *Proc. TRECVID 2012 Workshop*. Nov. 2012, Gaithersburg, MD, USA.

[11] A. Moumtzidou, P. Sidiropoulos, S. Vrochidis, N. Gkalelis, S. Nikolopoulos, V. Mezaris, I. Kompatsiaris, and I. Patras, "ITI-CERTH participation to TRECVID 2011," in *Proc. TRECVID 2011 Workshop*, Dec. 2011.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[13] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Linear subclass support vector machines," *IEEE Signal Process. Lett.*, vol. 19, no. 9, pp. 575–578, Sept. 2012.