IEEE International Conference on Image Processing (ICIP 2013), Melbourne, Australia, September 2013

VIDEO EVENT RECOUNTING USING MIXTURE SUBCLASS DISCRIMINANT ANALYSIS

Nikolaos Gkalelis [1,2], *Vasileios Mezaris* [1], *Ioannis Kompatsiaris* [1], *Tania Stathaki* [2]

[1] Information Technologies Institute / CERTH, Thermi 57001, Greece
[2] Electrical and Electronic Engineering Dept., Imperial College London, SW7 2AZ, UK

ABSTRACT

In this paper, a new feature selection method is used, in combination with a semantic model vector video representation, in order to enumerate the key semantic evidences of an event in a video signal. In particular, a set of semantic concept detectors is firstly used for estimating a model vector for each video signal, where each element of the model vector denotes the degree of confidence that the respective concept is depicted in the video. Then, a novel feature selection method is learned for each event of interest. This method is based on exploiting the first two eigenvectors derived using the eigenvalue formulation of the mixture subclass discriminant analysis. Subsequently, given a video-event pair, the proposed method jointly evaluates the significance of each concept for the detection of the given event and the degree of confidence with which this concept is detected in the given video, in order to decide which concepts provide the strongest evidence in support of the provided video-event link. Experimental results using a video collection of TRECVID demonstrate the effectiveness of the proposed video event recounting method.

Index Terms— Video event recounting, event detection, concept detection, semantic model vector, feature selection, mixture subclass discriminant analysis.

1. INTRODUCTION

High-level events can be conceived as dynamic objects that pace our everyday activities and provide the basis for structuring our memories [1]. For this reason, it is generally expected that event understanding technologies can contribute to effective organization of multimedia content and help in providing human comprehensible descriptions of this content to human users [2, 3].

During the past few years there has been a surge of research in the area of high-level event detection in video signals [4, 5, 6]. Video event detection algorithms typically derive one or more low-level features and then combine them using some fusion strategy. Recently, some researchers exploit a semantic model vector as a feature representation of high-level events, aiming at a better event detection performance (e.g., [7, 8, 9]). The inspiration behind this modelling approach is that high-level events can be better recognized by looking at their constituting semantic entities. However, another significant advantage of using a semantic model vector approach is that the derived video representation can serve as the first step towards automatic concept-based textual description of the video content and particularly of the visual evidence that supports the establishment of the video-event link [2, 3]. The latter is the goal of the emerging area of multimedia event recounting (MER) [10].

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

In MER, given a video that is said to belong to the target event class, the objective is to develop an event recounting algorithm (event recounter) that can produce a video event recounting document (VERD), describing in human-comprehensible format the key semantic entities that are depicted in this video and support the premise that the video belongs to the said event class. From the above definition we can make the following observations: a) A close relationship between event detection and recounting exists. For instance, the event recounter is usually applied on positive target event videos identified using an event detector (either a fully automatic one, or one involving additional manual inspection for filtering out the false alarm videos). b) Ideally, the VERD should be expressive enough so that a human subject could match it with the video clip it is derived from, as well as understand to which event class this VERD refers to [10]. c) Intuitively, a basic VERD type is one that contains a ranked list of concepts that are depicted in the video, e.g., "kicking", "ball" and "running" for a video presenting a soccer event. However, additional processing may be used to produce more advanced VERD types, e.g., consisting of higher complexity linguistic and/or non-linguistic entities, such as clauses, sentences, sounds, pictures, etc.

Only a limited number of works on video event recounting have been reported so far, mostly related to the pilot MER task of the TRECVID 2012 benchmark evaluation [10]. For instance, in [11], a large number of visual and audio semantic concept detectors are used to generate a concept-based representation of video segments, and the concepts detected with a high degree of confidence (DoC) are used to generate a single sentence for each video segment. Similarly, in [12], a semantic-based representation of the video is retrieved using several different technologies for the analysis of video content (keyword detection using automatic speech and optical character recognition; audio and visual concept detection at video- and video segment-level, etc.), and a VERD is delivered for each video exploiting the most relevant concepts according to their DoC values.

In this paper, a semantic-based approach is also used, however, instead of directly using the "raw" concept DoCs, a new feature selection approach is proposed that incorporates class separability criteria to select the most discriminant concepts regarding the target event. In particular, the linear discriminant analysis-based (LDA-based) feature selection method presented in [13] is extended using MSDA [14, 15] to handle nonlinearities in the data. The evaluation of the proposed method in the TRECVID MED 2010 dataset shows a significant performance gain over methods using directly the concept responses, or LDA-based selection of concepts.

The paper is organized as follows: A formulation of the video event recounting problem is provided in section 2. The model vector approach for representing videos is outlined in section 3 and the proposed video event recounting method is presented in section 4. Experimental results on TRECVID videos are given in section 5, while, conclusions and future work are discussed in section 6.

2. PROBLEM FORMULATION

Let $\mathcal{X} = \{(\mathbf{x}^p, y^p) \in \mathfrak{X} \times \{-1, 1\}\}$ be an annotated dataset representing videos belonging to the target event class ($y^p = 1$) or to the "rest of the world" event class ($y^p = -1$). Here, a model vector representation is used, i.e., $\mathbf{x}^p \in \mathfrak{X} \subset [0,1]^Q$ is the model vector derived from the p-th video. The κ-th element of $\mathbf{x}^p$ receives a value in $[0, 1]$ denoting the DoC that the semantic concept $c_\kappa \in \mathcal{C}$ is depicted in the video, where $\mathcal{C} = \{c_1, \ldots, c_Q\}$ is the set of concepts used in the model vector and Q is the total number of concepts. Our goal is to design a semantic concept selector (called hereafter event recounter) $r : \mathfrak{X} \rightarrow \{c_{n_1}, \ldots, c_{n_I}\}$, where n_i is the index of the i-th most relevant concept in the returned ordered list of concepts. That is, the event recounter receives as input the model vector representation of a video depicting the target event and provides as output a ranked list consisting of the $I \ll Q$ key semantic concepts that are depicted in this video and explain why the video relates to this event.

3. SEMANTIC MODEL VECTORS

A video signal is represented using a semantic model vector following a strategy similar to the one described in [9], as briefly explained in the next paragraphs.

The video signal is decoded and represented with a sequence of uniformly extracted keyframes. A two level spatial pyramid decomposition scheme, with an 1×3 cell division at the second level, is then combined with the dense sampling strategy and the oponentSIFT interest point descriptor in order to derive a set of 384-dimensional feature vectors for each pyramid cell [16]. Subsequently, a Bag-of-Words (BoW) model of 1000 visual words per pyramid cell is derived using the k-means algorithm, and a soft assignment procedure [17] is used for representing the overall keyframe with a 4000-dimensional BoW feature vector.

Then, a set of Q pre-trained concept detectors, $\mathcal{G} = \{g_\kappa : \mathbb{R}^{4000} \rightarrow [0, 1] | \kappa = 1, \ldots, Q\}$, is utilized to provide an intermediate level representation of a video keyframe based on Q semantic concepts [7, 8]. The concept detector g_κ is designed using a linear SVM and an appropriate training set referring to the κ-th semantic concept (this training set is independent from the set used for training and evaluating the event detectors and recounters). To this end, the t-th keyframe of the p-th video in the database is associated with the model vector $\mathbf{x}^{p,t} = [x^{p,1,t}, \ldots, x^{p,Q,t}]^T$, where, $x^{p,\kappa,t}$ is the response of the concept detector g_κ expressing the DoC that the κ-th concept is depicted in the keyframe.

In order to derive a model vector representation at video level, the model vectors of the individual keyframes referring to it are averaged. For instance, the model vector referring to the p-th video is computed using $\mathbf{x}^p = \sum_{t=1}^{T_p} \mathbf{x}^{p,t}$, where T_p is the length of the p-th video in keyframes.

4. EVENT RECOUNTING USING MSDA

The key semantic concepts describing the event depicted in the video signal can be retrieved by combining the model vector representation described above with an appropriate feature selection method. A feature selection method often consists of a selection criterion and a search strategy. It has been recently shown, that scatter matrix-based selection criteria such as the one used in linear discriminant analysis (LDA) offer competitive performance in comparison to the popular SVM-based algorithms [18].

In [13] the direct exploitation of LDA for feature selection is examined, yielding promising results. A training set partitioned to K classes $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$ (where in [13] each class corresponds to face images of a particular person) is exploited for the computation of the LDA transformation matrix $\mathbf{W} \in \mathbb{R}^{Q \times D}$, $D \ll Q$. Subsequently, the U first columns of $\mathbf{W}$ are utilized to provide a weight vector $\mathbf{v}$, whose κ-th component is the average of the absolute values of the respective components in the selected columns of $\mathbf{W}$, i.e., $v_\kappa = \sum_{i=1}^{U} |w_i^\kappa|$, where $|w_i^\kappa|$ is the absolute value of the κ-th element of the i-th column of $\mathbf{W}$. The I larger components of $\mathbf{v}$ are then selected to form a set $\mathcal{V} = \{\bar{v}_1, \ldots, \bar{v}_I\}$, i.e., $\bar{v}_\kappa$ is the κ-th largest component of $\mathbf{v}$, and then are utilized to design the feature selector

$$f_{lda} = \mathbf{a} \circledast \mathbf{x}, \tag{1}$$

where the operator $\circledast$ is used to denote element-wise vector multiplication. In (1), $\mathbf{a} \in \{0, 1\}^Q, \sum_{\kappa=1}^{Q} a_\kappa = I$ is a binary-valued indicator vector used to select the desired features from $\mathbf{x}$, whose κ-th element is $a_\kappa = 1$ if $v_\kappa \in \mathcal{V}$ and $a_\kappa = 0$ otherwise. The strength of this method is its computational efficiency and simplicity, as instead of using a time consuming search strategy a simple eigenvalue problem is solved. However, this method suffers from the nonlinearity problem of LDA, i.e., it faces difficulties to account for nonlinearities in the dataset. For instance in the case of event detection, which is a two-class problem, LDA can provide only one eigenvector which is not sufficient to capture such nonlinearities.

Inspired from the above method, we propose the use of MSDA [14] to build an efficient feature selection method. During the training stage, the iterative algorithm proposed in [14] is applied to derive a subclass partition of the data $\{\mathcal{X}_{1,1}, \ldots, \mathcal{X}_{K,H_K}\}$, where $\mathcal{X}_{i,j}$ denotes the j-th subclass of i-th class and H_i is the number of subclasses of class i. Subsequently, the transformation matrix $\mathbf{\Psi}$ is identified that maximizes the following objective function

$$J_{lda}(\mathbf{\Psi}) = \mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_{bsb} \mathbf{\Psi}) / \mathrm{tr}(\mathbf{\Psi}^T \breve{\mathbf{\Sigma}}_{\mathbf{X}} \mathbf{\Psi}), \tag{2}$$

where, $\breve{\mathbf{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{bsb} + \mathbf{S}_{ws}$, $\mathbf{S}_{bsb} = \sum_{i=1}^{K-1} \sum_{j=1}^{H_i} \sum_{k=i+1}^{K} \sum_{l=1}^{H_k} \hat{p}_{i,j} \hat{p}_{k,l} (\hat{\mathbf{m}}_{i,j} - \hat{\mathbf{m}}_{k,l})(\hat{\mathbf{m}}_{i,j} - \hat{\mathbf{m}}_{k,l})^T$ is the inter-between-subclass scatter matrix, $\mathbf{S}_{ws} = \sum_{i=1}^{K} \sum_{j=1}^{H_i} \hat{p}_{i,j} \hat{\mathbf{\Sigma}}_{i,j}$ is the within-class scatter matrix, H_i denotes the number of subclasses of the i-th class, and $\hat{p}_{i,j}$, $\hat{\mathbf{m}}_{i,j}$, $\hat{\mathbf{\Sigma}}_{i,j}$ are the estimated prior, sample mean and sample covariance matrix of the j-th subclass of class i. This optimization problem turns out to be equivalent to the generalized eigenvalue decomposition $\mathbf{S}_{bsb} \mathbf{\Psi} = \mathbf{S}_{ws} \mathbf{\Psi} \mathbf{\Lambda}$, where the columns of the transformation matrix, $\boldsymbol{\psi}_i$, $i = 1, \ldots, D$, are the generalized eigenvectors corresponding to the largest generalized eigenvalues in the diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$, and $\lambda_1 \geq \cdots \geq \lambda_D$. The first Ω columns of $\mathbf{\Psi}$ are then used to derive a weight vector $\boldsymbol{\phi} \in \mathbb{R}^Q$ whose κ-th element is computed using

$$\phi_\kappa = \max(|\psi_1^\kappa|, \ldots, |\psi_\Omega^\kappa|), \tag{3}$$

where, $|\psi_i^\kappa|$ is the absolute value of the κ-th element of the i-th column of $\mathbf{\Psi}$. Then, during testing, $\boldsymbol{\phi}$ can be utilized to design a feature selector similar to (1) (i.e., $\boldsymbol{\phi}$ could be used for deriving a binary-valued indicator vector $\boldsymbol{\alpha}$, to replace $\boldsymbol{\alpha}$ in (1)). However, in order to fully take advantage of the model vector video representation in building a recounting function, we further modify (1) so that $\boldsymbol{\phi}$ is directly used in it,

$$r_{msda} = \mathrm{sort}(\boldsymbol{\phi} \circledast \mathbf{x}, I), \tag{4}$$

where the $\mathrm{sort}(\mathbf{x}, I)$ vector operator returns the I largest values of $\mathbf{x}$ in descending order, along with the respective indices. In (4), the DoC concerning the presence of a concept in a video is weighted

with a significance value (computed according to (3)) that denotes the importance of this concept for the detection of the target event. With this modification the DoC that a concept is depicted in the particular video is also taken into account as opposed to using directly the feature selection approach of (1) [13]. Moreover, each concept is associated with a DoC regarding the target event, which allows the creation of a ranked list of concepts that best describe jointly the video and the event of interest depicted in it.

5. EXPERIMENTAL EVALUATION

5.1. Event recounters for comparison

During the evaluation procedure, the proposed recounter (4) is compared against two other event recounting approaches:

1) *Input space recounter*: the original concept detector responses can be used for selecting the highest-ranking concepts,

$$r_{in} = \text{sort}(\mathbf{x}, I) \tag{5}$$

2) *LDA-based recounter*: similar to the proposed MSDA-based method, an LDA-based recounter can be designed by modifying the feature selection approach in (1) [13] as follows:

$$r_{lda} = \text{sort}(\mathbf{v} \circledast \mathbf{x}, I) \tag{6}$$

5.2. Evaluation experiments and measures

The ability of an event recounter to generate a good VERD is evaluated according to the methodology of [10].

5.2.1. Verd-to-event experiment

In this experiment a set of unlabelled VERDs (which belong to E different events and were generated by O different recounters) are provided to S judges, and the task of each judge is to classify each VERD to one of the target events. Let $N_{o,e}$ denote the number of VERDs that belong to the e-th target event and were generated by the o-th recounter, and $\breve{N}^s_{o,e}$ ($\leq N_{o,e}$) the number of the above VERDs that are correctly classified to the e-th event by the s-th judge. Based on [10] the following measures are defined:

1) The verd-to-event correct classification rate for the o-th recounter and the e-th target event ($R^{o,e}_E$) is defined as the fraction of judgements of the s-th judge that correctly identified the target event, averaged across all judges: $R^{o,e}_E = \sum_{s=1}^{S} \breve{N}^s_{o,e}/(S \cdot N_{o,e})$.

2) The verd-to-event average correct classification rate of the o-th recounter (R^o_E) is defined as the average of $R^{o,e}_E$ across all events: $R^o_E = \sum_{e=1}^{E} R^{o,e}_E/E$.

5.2.2. Verd-to-clip experiment

In this experiment, for each event and each recounter a set of video clips and their associated VERDs are provided to S judges. That is, in total $O \cdot E$ sets are used: $(o, e) \in \{1, \ldots, O\} \times \{1, \ldots, E\}$; within each set, the clips belong to the same event and their VERDs have been generated by the same recounter. No information is provided to the judges regarding which recounter was used and to which event the VERDs refer to. The task of the judges is, within each set separately, to link the clips with their associated VERDs. This task examines how characteristic of each video the VERD descriptors are.

Let $M_{o,e}$ denote the number of VERDs that belong in the set associated with the o-th recounter and the e-th event class, and $\breve{M}^s_{o,e}$ ($\leq M_{o,e}$) the number of correct VERD-clip pairs provided by the s-th judge. Following [10] the following measures are used:

1) The verd-to-clip correct classification rate concerning the o-th recounter and the e-th target event ($R^{o,e}_C$) is defined as the fraction of valid VERD-clip pairs provided by the s-th judge, averaged across all judges: $R^{o,e}_C = \sum_{s=1}^{S} \breve{M}^s_{o,e}/(S \cdot M_{o,e})$.

2) The verd-to-clip average correct classification rate of the o-th recounter (R^o_C) is defined as the average of $R^{o,e}_C$ across all events: $R^o_C = \sum_{e=1}^{E} R^{o,e}_C/E$.

5.3. Datasets

Initially, the TRECVID SIN 2012 dataset is used to derive one concept detector for each of the $Q = 346$ SIN 2012 task concepts, as explained in section 3. Then, portions of the TRECVID MED 2010 dataset are used for training the event detectors and recounters, and evaluating the latter. The MED 2010 dataset consists of 1745 development and 1742 evaluation videos, belonging to one of three target events, namely, "assembling a shelter", "batting a run in" and "making a cake", or to the "rest-of-world" event class. Following the procedure described in section 3, and using the aforementioned 346 concept detectors, a model vector is estimated for each video.

Using the model vectors as features, one event detector is learned for each event on the development part of the MED 2010 dataset (event detectors are learned according to a variant of [9]). The trained event detectors are then used to automatically associate each video in the evaluation part of the MED 2010 dataset with a DoC for each event (i.e., they automatically detect videos belonging to the event). Thus, a ranked list of the videos is created for each event.

For evaluating the recounting method of section 4, three sets of video clips are then formed by selecting the first top 20 positive event clips from each ranked list using the ground-truth video annotation. Subsequently, the three recounters, r_{in}, r_{lda} and r_{msda}, are used to provide a VERD for each of the selected clips. The number of recounting concepts for each recounter and the number of selected columns in (3) for r_{msda} are set to $I = 15$ and $\Omega = 2$ respectively. In this way, 9 sets of VERD-clip pairs are created, where all the VERD-clip pairs within a set refer to a distinct "event-recounter" combination. The 180 VERDs created using the above procedure (20 VERDs per set) are used for the verd-to-event experiment. Additionally, 5 VERD-clip pairs are randomly selected from each set of 20 in order to form 9 evaluation sets for the verd-to-clip experiment (45 VERD-clip pairs in total).

5.4. Experimental results

In Figure 1 we present one video recounting example for each of the three events, along with the nine VERDs (three for each video) generated by the application of the three recounters (r_{in}, r_{lda} and r_{msda}) to the videos. In this example, we observe that the proposed recounter outperforms the other two recounters in terms of both number of correct concept detections as well as ordering of the correct concepts in the list. Apart from the improved concept selection performance, we also see that the proposed recounter effectively discards or ranks very low concepts such as "eucaryotic organism" and "primate", which are very generic and thus not suitable for describing the content of the video and the event taking place.

For the systematic evaluation of the recounters, experiments are carried out according to the experimentation methodology of section 5.2, using the datasets of section 5.3 and the judgement responses

Fig. 1. One video recounting example for each of the three events (from top to bottom): "assembling a shelter", "batting a run in" and "making a cake". Each row contains two keyframes of the video along with the responses of r_{in}, r_{lda} and r_{msda}. For each recounter response the top ten ranked concepts are shown; correctly identified concepts are presented with bold fonts, while wrong or very generic concepts with red italic or red normal font respectively.

Table 1. Results for the verd-to-event experiment.

Event	r_{in}	r_{lda}	r_{msda}
Assembl. shelter	0.800	0.850	**0.890**
Batting run in	0.490	0.870	**0.980**
Making cake	0.880	0.940	**0.960**
Average	0.723	0.887	**0.943**

Table 2. Results for the verd-to-clip experiment.

Event	r_{in}	r_{lda}	r_{msda}
Assembl. shelter	0.280	0.280	**0.440**
Batting run in	0.200	0.160	**0.240**
Making cake	**0.280**	**0.280**	0.120
Average	0.253	0.240	**0.267**

from five human judges ($S = 5$). The evaluation results are presented in Tables 5.4 and 5.4 for the verd-to-event and the verd-to-clip experiment respectively. We observe that:

a) For the verd-to-event experiment (Table 5.4), the exploitation of a feature selection process to build the event recounter (r_{lda} or r_{msda}) provides a significant improvement on the quality of the derived VERDs over the direct use of the concept DoCs (r_{in}). For instance, for the event "batting a run in", the 0.49 of $R_E^{o,e}$ achieved with r_{in} is increased to 0.98 utilizing r_{msda}. Using r_{msda}, a significant average performance boost of approximately 30.4% and 6.4% is observed over r_{in} and r_{lda} respectively.

b) For the verd-to-clip experiment (Table 5.4), the proposed recounter provides a significant performance gain over the two other methods for two of the three considered events. Overall, achieves an average performance boost of 5.2% and 11.1% over r_{in} and r_{lda} respectively.

c) The verd-to-clip task seems to be more challenging than the verd-to-event one as indicated by the results. This can probably be attributed to the number of concepts included in each VERD in our experiment ($I = 15$), and to shortcomings of the overall set of 346 concepts in fully describing the video content. Nevertheless, the proposed method still provides improved performance in this experiment as well, by effectively removing highly-ranked erroneous or very generic concepts.

6. CONCLUSIONS AND FUTURE WORK

A novel event recounting method was presented that exploits MSDA and a semantic video representation in order to extract the key visual concepts of the depicted event. The evaluation of the proposed method using the TRECVID MED 2010 video collection and five human experts demonstrated a significant improvement over methods that exploit conventional linear discriminant analysis or use directly raw semantic concept detector responses.

Interesting future research directions include the exploitation of language templates to provide richer event descriptions and the investigation of methods to select the optimal number of semantic concepts for event description (instead of $I = \text{const}$).

7. REFERENCES

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

[2] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Automatic event-based indexing of multimedia content using a joint content-event model," in *ACM Multimedia 2010, (EiMM)*, Firenze, Italy, Oct. 2010.

[3] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "A joint content-event model for event-centric multimedia indexing," in *Proc. 2010 IEEE 4th Int. Conf. on Semantic Computing*, Pittsburgh, PA, USA, Sept. 2010, pp. 79–84.

[4] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimed. Info. Retr.*, Nov. 2012.

[5] Y. Kamishima, N. Inoue, K. Shinoda, and S. Sato, "Multimedia event detection using GMM supervectors and SVMs," in *Proc. IEEE Int. Conf. Image Processing*, Orlando, Florida, USA, Sept./Oct. 2012, pp. 3089–3092.

[6] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Estimation and representation of accumulated motion characteristics for semantic event detection," in *Proc. IEEE Int. Conf. Image Processing, (MIR)*, San Diego, California, USA, Oct. 2008, pp. 3089–3092.

[7] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "High-level event detection in video exploiting discriminant concepts," in *Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011)*, Madrid, Spain, June 2011, pp. 85–90.

[8] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[9] A. Moumtzidou, N. Gkalelis, and P. Sidiropoulos et al., "ITI-CERTH participation to TRECVID 2012," in *Proc. TRECVID 2012 Workshop*.

[10] P. Over, G. Awad, and M. Michel et al., "TRECVID 2012 - an introduction to the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.

[11] L. Cao, S.-F. Chang, and N. Codella et al., "IBM Research and Columbia University TRECVID-2012 multimedia event detection (MED), multimedia event recounting (MER), and semantic indexing (SIN) systems," in *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.

[12] P. Natarajan et al., "BBN VISER TRECVID 2012 multimedia event detection and multimedia event recounting systems," in *Proc. TRECVID 2012 Workshop*, Gaithersburg, MD, USA, Nov. 2012.

[13] F. Song, D. Mei, and H. Li, "Feature selection based on linear discriminant analysis," in *Proc. 2010 Int. Conf. Intelligent System Design and Engineering Application*, Changsha, China, Nov./Dec. 2010, vol. 1, pp. 746–749.

[14] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Mixture subclass discriminant analysis," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 319–332, May 2011.

[15] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 24, no. 1, pp. 8–21, Jan. 2013.

[16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sept. 2010.

[17] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Sept. 2010.

[18] L. Wang, "Feature selection with kernel class separability," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1534–1546, Sept. 2008.