論文 / 著書情報
Article / Book Information

Title	Event detection in consumer videos using GMM supervectors and SVMs
Authors	Yusuke Kamishima, Nakamasa Inoue, Koichi Shinoda
Citation(English)	EURASIP Journal on Image and Video Processing, vol. 2013:51, , pp. 1-13
Pub. date	2013, 9
Creative Commons	次ページ参照

License

Creative Commons：表示

EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Event detection in consumer videos using GMM supervectors and SVMs

Yusuke Kamishima, Nakamasa Inoue* and Koichi Shinoda

Abstract

In large-scale multimedia event detection, complex target events are extracted from a large set of consumer-generated web videos taken in unconstrained environments. We devised a multimedia event detection method based on Gaussian mixture model (GMM) supervectors and support vector machines. A GMM supervector consists of the parameters of a GMM for the distribution of low-level features extracted from a video clip. A GMM is regarded as an extension of the bag-of-words framework to a probabilistic framework, and thus, it can be expected to be robust against the data insufficiency problem. We also propose a camera motion cancelled feature, which is a spatio-temporal feature robust against camera motions found in consumer-generated web videos. By combining these methods with the existing features, we aim to construct a high-performance event detection system. The effectiveness of our method is evaluated using TRECVID MED task benchmark.

Keywords: Multimedia event detection; Feature extraction; GMM supervector; Support vector machines; Camera motion cancelled features

1 Introduction

The amount of consumer-generated web videos we can access over the Internet has been rapidly increasing. For example in Youtube, more than 72 h of video are uploaded per a minute. Since such videos often do not have text tags, there has been a strong demand for automatic video retrieval systems based on video contents. In particular, detecting *events* depicted in a video enables us to get significant information. Here, events are characterized by the combination of several *concepts* such as objects, scenes, and human motions. For example, an event *birthday party* consists of concepts such as *cake*, *indoor*, *singing*, and *people*[a].

Most studies for event detection have been aimed at identifying events in professionally produced videos such as sports [1] and movies [2], or in surveillance videos [3]. These studies used event-specific methods which rely heavily on the spatial-temporal structures of the target events. By Assfalg et al. [1], for example, three types of highlights in soccer games such as *penalty kick* were modelled by three-state HMMs, using constant camera motions and the location of players as features. Li et al. [2] detected dialog events in movie videos by shot clustering and audio analysis. Adam et al. [3] modelled the stream of people in surveillance videos from optical flows to detect unusual events such as *running in the mall*. While these methods can be applied when the target events are specified and the spatial-temporal structures of the events are always stable, it is difficult to apply them to general events appearing in consumer-generated videos due to two major reasons. First is that general events widely vary and the definitions of them are not always clear. Second is that consumer-generated videos do not have stable spatial-temporal characteristics since they are taken by amateurs from different points of view and often include unsettled camera motions or haphazard editing.

Our aim is to detect an event consisting of multiple concepts from a large amount of consumer-generated videos. One attempt at tackling such complicated event detection is the TRECVID [4] multimedia event detection (MED) task [5]. In this task, several events are described in text using more than one concept. Some examples are shown in Figure 1. The MED dataset for this task consists of a large amount of clips, which includes various types of videos such as home videos and demonstration videos.

*Correspondence: inoue@ks.cs.titech.ac.jp
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Figure 1 **Examples of event clips in TRECVID2010 and TRECVID2011 MED dataset.**

For example, the number of clips in the 2011 task was 44,904.

Many studies for this kind of task have been based on the bag-of-words (BoW) framework [6-9]. In these methods, low-level features are converted into a fixed-length BoW histogram [10,11], based on the assignment of each feature to one code in a codebook. This histogram is used as the input to an support vector machine (SVM). However, since this BoW method uses hard-assignment (i.e., a feature vector is assigned only to the nearest codeword), quantization errors often degrade the detection performance.

Some event detection methods used not only low-level features and BoW methods but also high-level features such as a semantic concept model, which models the relationship between events and concepts [6,12]. Jiang et al. [6] applied the domain-adaptive semantic diffusion (DASD) method [13] to MED, where the labels of the concepts related to the target events were manually annotated to positive clips for MED events and the relationship was modelled using a weighted graph. The weights were estimated using training samples and automatically adapted to testing data. While such a method can give us the contextual information and fill the semantic gap between low-level features and detection results, it does not have extensibility since we do not know what/how many concepts are needed for such kind of event modelling. To address this problem, Ma et al. [12] proposed event detection using a classifier-specific intermediate representation, where the correlation of an event and concepts is automatically learned and utilized for event detection. It used both event-based and concept-based datasets. Although the use of other datasets is cost efficient, the differences in the characteristics of datasets such as video quality and semantic contents make it difficult to learn the relationship between an event and concepts precisely.

To address this *data insufficiency* problem, the combination of GMM supervectors and SVMs (GSSVM) was recently proposed [14,15] for detecting concepts in video shots. In this method, a Gaussian mixture model (GMM) represents the distribution of local features extracted from

a video shot and a GMM supervector is created from the GMM parameters. The GMM supervector for each shot is used as the input to an SVM.

It can be regarded as one of the recent high-dimensional image representations such as Fisher vector [16] and super-vector [17]. Fisher vector represents an image by the gradient of the log-likelihoods obtained using a GMM. Supervector, which uses the first order differences between low-level features and visual words, is a simplification of the Fisher vector. For a video representation, since the number of low-level features varies depending on the length of a video, GMM supervector which is derived based on maximum *a posteriori* (MAP) estimation is expected to be effective and be robust to such variation.

We consider to apply this GSSVM method to event detection. In event detection, data collection presents a difficulty. Since the contents of consumer-generated videos widely vary, the amount of positive samples tends to be smaller than the negative samples. In the TRECVID2011 MED training dataset, the number of positive clips for each event (mean, 139) is relatively smaller than that of negative clips (mean, 12,944). In such a case, since it is difficult to learn an event precisely with a hard-assignment like BoW, GMM is expected to be effective. On the other hand, GSSVM might not fill the semantic gap between low-level features and detection results while the semantic event models such as [6,12] might, if we have an efficient amount of concept data. However, collecting concept data manually costs a lot and may be endless since target events vary. In the current situation, where we do not have sufficient amounts of data, we need an event model such as GSSVM, whose parameters can be robustly estimated even when the amount of training data is small. Another difficulty in event detection is that consumer-generated videos do not have the spatial-temporal structure specific to each event due to the variety of video characteristics such as durations, camera motions, and editing. Thus, our method does not explicitly utilize the spatial-temporal structure of each event, which is difficult to model from consumer-generated videos. We expect that, even without them, the combination of GMMs and SVMs will have high detection performance.

Thus, in this paper, we propose an event detection method based on GSSVM [18] and aim at constructing an event detector robust against the data insufficiency. Event detection using GMMs of scale-invariant feature transform (SIFT) features [19] from unconstrained news shots was proposed in [20]. To the best of our knowledge, we are the first to apply this framework to general event detection in consumer-generated videos.

Another important factor in event detection is what low-level features should be used. Many local features have been proposed for image and video recognition. SIFT [19] has been effective in several image (e.g., [21]) and video application (e.g., [20]). SURF [22] needs several times less computation than SIFT. MFCC features, which have been often used for speech recognition, were proven to be effective for video recognition. In video recognition, the combination of more than one feature has been shown to be more effective than each feature alone [23]. In addition to visual and audio features, motion features such as space time interest points (STIP) [24] have often been used for event detection. STIP, originally proposed for action recognition, are regions detected using a Harris 3D detector that have significant local spatial-temporal changes. Each interest point is described with either of two types of descriptors: histograms of oriented gradients (HOG) or histograms of optical flows (HOF). Wang et al. [25] extended this feature and found the dense sampling of these two descriptors was also effective for action recognition. Kläser et al. [26] proposed 3DHOG, which is an extension of HOG to spatial-temporal space. Chen and Hauptmann [27] proposed motion SIFT (MoSIFT), which combines information from SIFT and optical flow. Such spatial-temporal features were introduced to event detection in [6,8], and they improved the detection rate when combined with visual and audio features.

However, these motion features may be affected by camera motions. When the camera has motions such as translation and zoom, the motions of the foreground objects are different from their true motions and accordingly, spatial-temporal features do not represent the true foreground motions. Some previous studies for action recognition have addressed this problem. Foreground motions, background motions, and camera motions are decomposed based on pixel trajectories in [28]. Motion boundary histogram (MBH) [29] based on the derivatives of optical flows around the trajectory of pixels was also proposed as a robust feature against camera motion. Mikolajczyk and Uemura's work [30] used matching of feature points for camera motion estimation. Ikizler-Cinbis and Sclaroff [31] used the segmentation of frame images for separation of foreground objects and background planes, and then, camera motions are estimated from background planes. However, since these methods are computationally expensive and the target events are limited to simple ones such as 'walking' and 'diving,' it is difficult to apply them to large-scale event detection.

Li et al. [32] categorized camera motions into four categories (static, pan, tilt, zoom). The categories are used as features combined with other types of features such as foreground object motions, background object motions, and scale of foreground objects. While it is a sophisticated algorithm dealing with camera motions, major foreground objects that have important meaning in multimedia events are expected to be detected by capturing

only translations since they are often in the center of a video with camera panning.

To overcome such problems, we propose camera motion cancelled features. They are extracted by estimating camera motion for each frame from optical flows and cancelling it before feature description. Optical flows are computed in the peripheral region in a frame image. Since no tracking process is needed, it is applicable to large-scale event detection. We fuse this feature with other existing low-level features, which have complementary information, in a late fusion framework.

The new contributions of this paper are (1) the application of our GMM supervector in [14,15] to MED and (2) the camera motion cancelled features. In [14,15], we have applied GMM supervector to video semantic indexing and have reported that GSSVM performs the best in the TRECVID 2011 semantic indexing (SIN) task, where word-level semantic concepts such as 'airplane' and 'sky' were targets. In this paper, we evaluate GSSVM using datasets from the TRECVID 2010 and 2011 MED tasks, where sentence-level events such as 'making sandwich' and 'batting in a run in a baseball game' were targets. Here, to detect such events, we train GSSVMs not only for appearance features such as SIFT as in our previous study [14,15,33] but also for STIP motion features and its camera-motion-cancelled version.

The rest of this paper is organized as follows. Section 2 explains the overview of our detection system. Section 3 explains the existing seven low-level features. Section 4 explains our proposed camera motion cancelled feature. Section 5 explains GSSVM. Section 6 describes the experimental conditions, results, and analysis. Finally, we conclude the paper in Section 7.

2 Overview of our event detection system

Figure 2 shows an overview of our system, consisting of three major phases: feature extraction, modelling, and detection. In the feature extraction phase, the video input is processed to extract eight types of low-level features. To increase the robustness against the camera motion, we introduce camera motion cancelled features (CC-DSTIP). In the modelling phase, a GMM supervector is created for each clip using the extracted low-level features. For visual features, the spatial pyramid matching is used. In the detection phase, SVM scores for each feature are fused by their weighted average. If the score is above a threshold, the clip is predicted to include the event, otherwise not.

3 Low-level features

Since videos have multi-modality, it is important to use multiple features to build a high-accuracy multimedia event detection system. We use eight types of complementary features: five visual features, one audio feature, and two motion features (STIP and CC-DSTIP). We explain about CC-DSTIP, our new feature, in the next section. To use the location information of feature vectors, spatial pyramid matching is applied to visual features. Principal components analysis (PCA) is applied to all features in order to reduce the number of dimension for saving computational costs in the training and detection steps.

3.1 Visual features

- SIFT with Harris-Affine region detector (*SIFT-Har*): Scale-invariant feature transform (SIFT) [19] has been effective in many researches of image and video analysis such as concept detection in video shots [14,15], since it is invariant to image scaling and illumination change. Before the computation of feature vectors, Harris-Affine detector [34], which is an extension of Harris corner detector, is applied to gray-scale frame image. Then, a 128-dimensional feature vector is computed from the intensity gradients for each region. Since extracting SIFT features from all the clips and all of the frames is computationally too expensive, features are extracted from one frame image every 2 s (similarly for SIFT-Hes, SURF, and HOG). The dimension is reduced to 32 by PCA.
- SIFT with Hessian-0Affine region detector (*SIFT-Hes*): We also use SIFT features extracted from the Hessian-Affine regions [34]. The Hessian-Affine region detector is often used to detect blobs and is known to be complementary to the Harris-Affine region detector. The combination of different detectors can improve a method's robustness to noise. PCA is applied to reduce the dimension from 128 to 32.
- SURF features (*SURF*): Speeded up robust features (SURF) [22], which are several times faster to extract than SIFT, are extracted using sums of 2D Haar Wavelet responses. It is often used for image matching. The dimension is reduced from 64 to 32 by PCA.
- HOG features with dense sampling (*HOG*): We also use histogram of oriented gradients (HOG) [35] features sampled densely from an image. The dense sampling of HOG has less computational cost than SIFT does. A vector consists of eight-bin histograms of gradients extracted from 2×2 blocks (32 dimensions). Different from keypoint-based features such as SIFT or SURF, dense sampling gives us a fixed number of features. PCA is applied keeping the number of dimensions the same.
- RGB-SIFT features with dense sampling (*RGB-SIFT*): Color information is often helpful for video analysis.

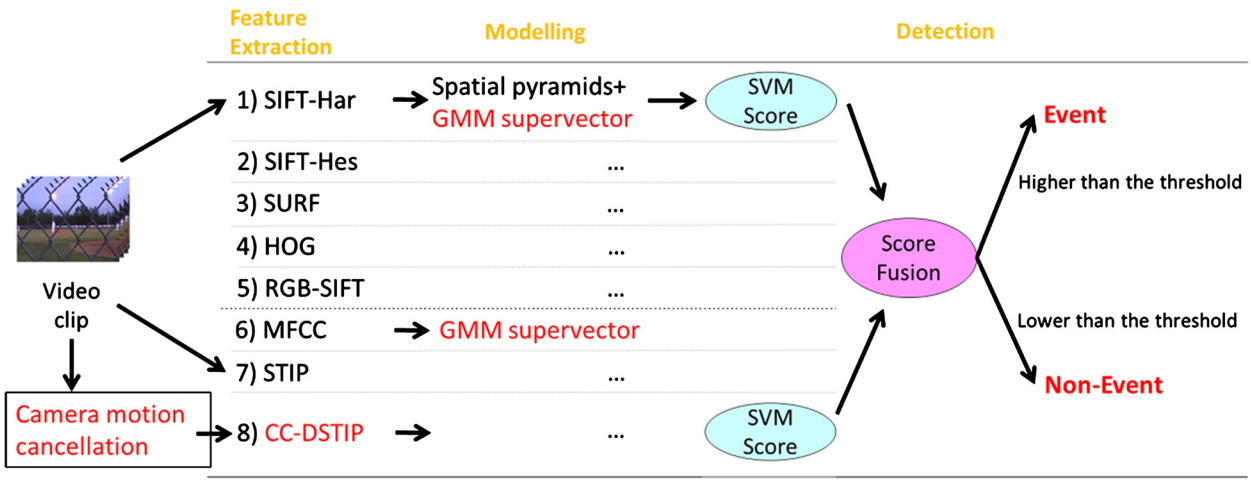

Figure 2 System overview. In the feature extraction phase, eight types of features are extracted for each clip. CC-DSTIP is our new feature. In the modelling phase, GMM supervectors corresponding to eight features are constructed. In the detection phase, the GMM supervector for each feature is used as the input to an SVM, and finally, SVM scores are fused by weighted average. If the fused detection score is higher than the threshold, the clip is detected as an event and otherwise not.

RGB-SIFT features [36] are the concatenated SIFT features extracted from each of RGB channel of an image. Features are sampled from one frame image every 6 s. PCA is applied to reduce the dimensionality from 384 to 64.

3.2 Audio feature

MFCC features (*MFCC*): Audio is an important clue when analyzing video content. We use Mel frequency cepstral coefficient (MFCC) features, which are often used in speech recognition. In addition, we use ΔMFCC, ΔΔMFCC, Δpower, and ΔΔpower. The dimension of a feature vector is 38. We compute the MFCC feature over a 24-ms time window with a 12-ms overlap. PCA is applied keeping the number of dimensions same.

3.3 Motion feature

Spatial-temporal features with Harris 3D detector (*STIP*): Space-time interest points (STIP) [24] are the region induced using Harris 3D detector, which is an extension of Harris-corner detector to three dimensions: the horizontal direction x, vertical direction y, and temporal direction t. Features extracted from STIP are expected to represent motions since the regions detected as STIP have significant spatial and temporal changes. We sample four-bin HOG features and five-bin histograms of optical flow (HOF) features from $n_x \times n_y \times n_t$ blocks in each STIP. In this work, we set $n_x = 3$, $n_y = 3$, and $n_t = 2$. Then, a 72-dimensional HOG feature and a 90-dimensional HOF feature are concatenated into a 162-dimensional vector (HOGHOF). The dimension is reduced to 64 with PCA.

3.4 Spatial pyramid matching

Spatial pyramid matching [37] enables us to use the location information of feature vectors. In this method, a fixed length vector, such as bag-of-words histogram or GMM supervector (Section 5), is constructed from low-level features in each of the divided regions of a video clip. Then, the fixed length vectors for all the regions in the clip are concatenated into one vector. We apply the pyramids with three levels (1×1, 2×2, and 3×1) like Figure 3 for five types of visual features. As a result, a vector of eight times the length of the original GMM supervector is computed for each clip.

4 Camera motion cancelled features

Although motion analysis is necessary for event detection, many clips contain camera motions, and thus, the motions of the foreground objects in such clips are different from their true motions.

For large-scale consumer-generated video archives, a simple and fast method is effective since camera motions in consumer-generated videos are often very simple compared to those in professional videos. Our proposed method uses voting of optical flows to capture camera motions. It captures horizontal and vertical slow panning, which is the most frequent camera motion in consumer-generated videos, to remove camera motions. While complicated camera motions such as zoom, tilt, and their combinations cannot be captured, major foreground objects that have important meaning in multimedia events are expected to be detected since they are often in the center of a video with camera panning.

Camera motion for each frame is estimated using optical flows. Since the center region of a frame tends to

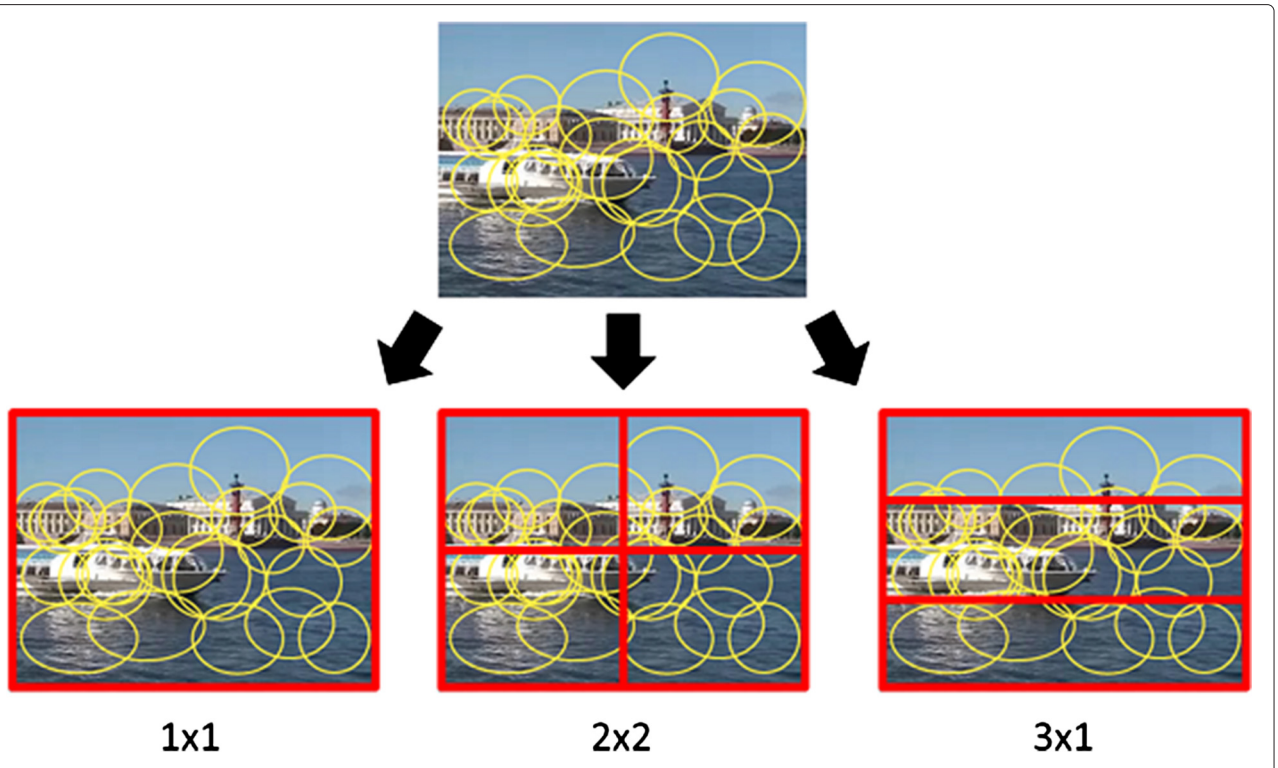

Figure 3 Three-level spatial pyramid.

include foreground objects, only the optical flows within the peripheral region are used for camera motion estimation. $R(\alpha W, \alpha H)$ denotes the peripheral region having αW width from the left and right of a video, and αH height from the top and bottom (Figure 4a). Here, W and H are the width and height of the video, respectively, and α is an experimentally decided parameter. Optical flows in $R(\alpha W, \alpha H)$ are computed every five pixels. The camera motion is decided by the voting of quantized optical flows in a frame image. We quantize each two-dimensional optical flow f_i $(i = 1, \ldots, N_f)$ as

$$q_{i,(u,v)} \in \{0,1\},\ q_{i,(u,v)} = 1 \iff f_i = (u,v), \tag{1}$$

where N_f is the number of optical flows in $R(\alpha W, \alpha H)$ and the integer values, u and v, are the flows for the

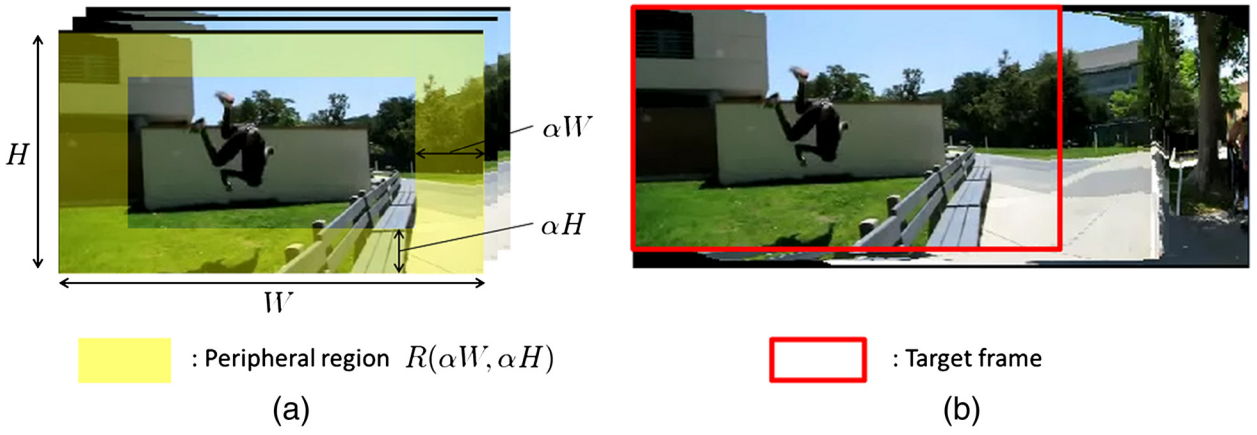

Figure 4 An example of camera motion cancellation. (a) shows the frames before cancellation and **(b)** shows the frames after cancellation. Optical flows are computed from the peripheral region $R(\alpha W, \alpha H)$ in **(a)** and HOGHOF features are described from the target frame in **(b)**.

horizontal and vertical direction, respectively. Then, the camera motion ω is given by

$$\omega = \begin{cases} arg\max_{(u,v)} \sum_{i=1}^{N_f} q_{i,(u,v)} & \text{if } \max_{(u,v)} \sum_{i=1}^{N_f} q_{i,(u,v)} \geq \epsilon N_f \\ (0,0) & \text{otherwise.} \end{cases} \tag{2}$$

where ϵ ($\leq$ 1) is the parameter to avoid the false estimation of camera motion. We move each frame image to the same direction with the same length as the camera motion ω and then extract spatial-temporal HOGHOF features with dense sampling [25] from the target frame (Figure 4b). The dimensionality of HOGHOF is reduced from 162 to 64 by PCA. We call this feature **CC-DSTIP** in the following sections.

5 GMM supervectors and SVMs

The combination of GMM supervectors and SVMs (GSSVM) was first proposed for speaker verification [38]. In [14,15,33], we have applied it to video semantic indexing and have reported that GSSVM performed the best in the TRECVID 2011 SIN task. However, since TRECVID SIN aims at detecting word-level semantic concepts such as 'airplane,' 'car,' 'sky,' and 'cityscape,' it has not been clear whether the GSSVM is effective for MED aiming at detecting sentence-level events consisting of multiple concepts and actions such as 'making a sandwich' and 'batting in a run in a baseball game.' Here, to detect such events, we train GSSVMs not only for appearance features such as SIFT as in our previous studies [14,15,33] but also for STIP motion features and its camera-motion-cancelled version.

We first make a GMM for a set of (a type of) feature vectors. Then, we construct a GMM supervector from each GMM by MAP adaptation. In the detection phase, we use it as an input for a SVM classifier. Finally, we fuse the outputs of the SVMs for the eight feature types and use the weighted average of them as the detection score. We explain each step in this section.

5.1 Gaussian mixture models

A GMM, whose probability density function is given by

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{3}$$

is used to model a video clip. Here, x is a feature vector for one of the low-level feature types, $\theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ is a set of GMM parameters, K is the number of Gaussian mixture components (vocabulary size), w_k is the weight for mixture component k, and $\mathcal{N}(x|\mu_k, \Sigma_k)$ is a Gaussian probability density function with a mean vector μ_k and a covariance matrix Σ_k for mixture component k given by

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^{\mathrm{T}} \Sigma_k^{-1} (x-\mu_k)}, \tag{4}$$

where d is the dimension of x.

5.2 MAP adaptation

The GMM parameters are estimated for each clip using the MAP criterion. This process is often called MAP adaptation [39]. In this adaptation, the parameters of a universal background model (UBM), which are estimated from all video clips using expectation maximization (EM) algorithm, are utilized as the prior distribution for Gaussian means. This adaptation is particularly effective when the amount of data available is small. Let $\theta^{(U)} = \{w_k^{(U)}, \mu_k^{(U)}, \Sigma_k^{(U)}\}_{k=1}^{K}$ be the parameter set of UBM U, where $w_k^{(U)}$ is the weight, $\mu_k^{(U)}$ is the mean vector, $\Sigma_k^{(U)}$ is the covariance matrix for the mixture component k of U, respectively.

Then, the MAP estimate $\hat{\mu}_k$ for the Gaussian mean is

$$\hat{\mu}_k = \frac{\tau \mu_k^{(U)} + \sum_{i=1}^{n} c_{ik} x_i}{\tau + \sum_{i=1}^{n} c_{ik}}, \tag{5}$$

$$c_{ik} = \frac{w_k^{(U)} \mathcal{N}(x_i|\mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k^{(U)} \mathcal{N}(x_i|\mu_k^{(U)}, \Sigma_k^{(U)})}, \tag{6}$$

where $X = \{x_i\}_{i=1}^{n}$ is a feature vector set of one of the eight feature types extracted from a video clip, c_{ik} is the contribution rate of x_i for the k-th Gaussian component (the posterior probability of x_i being at the k-th Gaussian component), and τ is a hyper-parameter which controls the weight of the prior against the maximum likelihood estimate.

5.3 GMM supervector

After MAP adaptation, a GMM supervector $\phi(X)$ is constructed for each video clip by concatenating the mean vectors of all the mixture components in the corresponding GMM as:

$$\phi(X) = (\tilde{\mu}_1^{\mathrm{T}} \tilde{\mu}_2^{\mathrm{T}} \dots \tilde{\mu}_K^{\mathrm{T}})^{\mathrm{T}}, \ \tilde{\mu}_k = \sqrt{w_k^{(U)}} (\Sigma_k^{(U)})^{-\frac{1}{2}} \hat{\mu}_k. \tag{7}$$

Here, each mean vector is normalized by its related weight and variance. This GMM supervector is then input to the support vector machine.

5.4 Detection using support vector machines

We use support vector machines (SVMs) with the following RBF-kernel for each of the low-level feature types to detect each event:

$$k(X_i, X_j) = \exp(-\gamma \|\phi(X_i) - \phi(X_j)\|_2^2), \tag{8}$$

where $\|x\|_2^2$ is the squared 2-norm of x, X_i, and X_j are sets of feature vectors and γ is an experimentally optimized control parameter. We set γ to the inverse of the average distance between two GMM supervectors. The SVM discriminative function is given by

$$f(X) = \sum_{l=1}^{L} a_l k(X, X_l) + b, \tag{9}$$

where X_l is the set of feature vectors corresponding to a training video clip. L is the number of the training video clips. a_l and b are the SVM parameters set during the training step. We train an SVM for each event and each feature.

5.5 Late fusion of features

The detection score for the event E is given by

$$s_E(X) = \sum_{\mathrm{F}} \beta_{E,\mathrm{F}} f_{E,\mathrm{F}}(X), \tag{10}$$

where $f_{E,\mathrm{F}}$ is the prediction score, which is the output of an SVM discriminative function for event E trained using feature type F $\in$ {SIFT-Har, SIFT-Hes, SURF, HOG, RGB-SIFT, MFCC, STIP, CC-DSTIP}, and $\beta_{E,\mathrm{F}}$ is the fusion weight for E and F. We decide $\beta_{E,\mathrm{F}}$ by twofold cross validation using training data. The scores are finally normalized into [0,1] domain using the sigmoid function.

6 Experiments

6.1 Conditions

We used the video dataset of the MED task in TRECVID2010 and TRECVID2011. The TRECVID2010 MED (MED10) dataset has 3,468 videos, of which 1,744 videos are for training and 1,724 are for testing. The target events are manually annotated and consist of 'Assembling a shelter,' 'Batting a run in,' and 'Making a cake.' The positive clips of each event amount to about 50 for training and 50 for testing. The TRECVID2011 MED (MED11) dataset has 44,904 videos, of which 13,083 videos are for training and 31,821 videos are for testing. Ten target events are listed on Figure 1. Each event has between 78-231 positive clips for both training and testing.

The evaluation criteria are the same in both TRECVID MED tasks. These criteria are mainly based on the missed detection rate (P_{MD}), false alarm rate (P_{FA}), and normalized detection cost (NDC). The NDC is a linear combination of the probabilities of two types of errors: P_{MD} and P_{FA}. NDC, P_{MD}, and P_{FA} are given by

$$\mathrm{NDC}(T) = w_{\mathrm{MD}} P_{\mathrm{MD}}(T) + w_{\mathrm{FA}} P_{\mathrm{FA}}(T), \tag{11}$$

$$P_{\mathrm{MD}}(T) = N_{\mathrm{MD}}(T)/N_{\mathrm{pos}}, \tag{12}$$

$$P_{\mathrm{FA}}(T) = N_{\mathrm{FA}}(T)/N_{\mathrm{neg}}, \tag{13}$$

where w_{MD} and w_{FA} are parameters which control the weights of the missed detection rate and false alarm rate, respectively. In TRECVID MED and this work, $w_{\mathrm{MD}} = 1.0, w_{\mathrm{FA}} = 12.49$. T is the detection threshold. N_{MD} is the number of positive clips with the score under the threshold T, and N_{FA} is the number of negative clips with score above T. N_{pos} and N_{neg} are the number of the positive clips and negative clips, respectively. Note that lower NDC indicates better performance. Further, we have another type of NDC, minimum NDC (MinNDC). This is the minimum of NDC over the detection threshold T, which is the value when the detection threshold T is optimized posteriorly as:

$$\mathrm{MinNDC} = \min_{T} \mathrm{NDC}(T). \tag{14}$$

The comparison of NDC and MinNDC gives us how close the detection threshold is to the optimized one under the NDC criteria and the potential performance when the threshold is appropriately decided.

SIFT, SURF, RGBSIFT, STIP, and MFCC features were extracted using the implementations in [22,24,36,40], and [41], respectively. HOG was extracted using our own implementation. The description of dense HOGHOF features in CC-DSTIP was also done using [24]. For the SVM training and prediction, libSVM [42] was used.

The number of Gaussian components K was selected from 256, 512, and 1,024 by twofold cross validation (CV) on the training set in which we obtained mean NDC of 0.743, 0.731, and 0.731, respectively. Since using a smaller number is computationally efficient, we selected $K = 512$ which performed as well as $K = 1,024$. The fusion weight $\beta_{E,F}$ and the detection threshold are selected by grid search with a step size of 0.01 with twofold CV. The hyper-parameter τ is set to 20.0, which is the default value of Hidden Markov Toolkit (HTK) [41]. In camera motion cancellation, parameters α and ϵ were set to 0.2 and 0.7, respectively, since our preliminary experiment with 20

training video clips showed that estimation of the camera motion is robust to these parameters when $0.1 \leq \alpha \leq 0.3$ and $0.5 \leq \epsilon \leq 0.9$.

6.2 Results

6.2.1 Comparison of GSSVM with bag-of-words and semantic concept model-based methods

We compared our method with the previous method proposed by Jiang et al. [6] which combined BoW and DASD. It should be noted that it had the best performance in the original TRECVID2010 MED competition. We show the result in Table 1. Since we used different features from theirs, it is difficult to directly compare the performance. Our method achieved mean MinNDC 0.558 when we used three features: SIFT-Hes, MFCC, and STIP. Their method had mean MinNDC 0.579 when they used not only the same three features but also another feature, SIFT with difference of Gaussian (SIFT-DoG in Table 1) [43], which is more likely to detect edges than the Harris-Affine detector and the Hessian-Affine detector. The performance of their method improved to mean MinNDC 0.565 when they additionally used earth mover's distance (EMD) [44] as a metric between two BoW histograms. The performance of our method was significantly better than these two results, and thus, the effectiveness of our method was confirmed. Furthermore, our performance was improved to mean MinNDC 0.510 when all eight types of low-level features in Section 3 and 4 were used.

SCV in Table 1 is our trial of semantic event model using concept detectors. We aggregated the detection score of 346 concepts defined in TRECVID2011 SIN task [45] and constructed a 346-dimensional score vector (SCV) for each of MED clips. The concept detectors were learned from GMM supervectors with HOG features using the SIN dataset. The score vectors were used as the input to an event SVM. As a result, the mean MinNDC of HOG-SCV (0.719) was much higher than that of the combination of HOG-GSSVM (0.614). It is because the number of concepts was not enough and most of the 346 concepts, which had been selected independently of the MED events, were not useful for event detection.

Table 1 Mean MinNDC on TRECVID2010 MED

Features - methods	Mean MinNDC
SIFT-DoG + SIFT-Hes + MFCC + STIP - BoW [6]	0.586
SIFT-DoG + SIFT-Hes + MFCC + STIP - BoW + DASD [6]	0.579
SIFT-DoG + SIFT-Hes + MFCC + STIP - BoW + DASD + EMD [6]	0.565
SIFT-Hes + MFCC + STIP - *GSSVM*	0.558
All 8 low-level features - *GSSVM*	0.510
HOG - SCV	0.719
HOG - *GSSVM*	0.614

Mean MinNDCs of the bag-of-words (BoW) and domain-adaptive semantic diffusion (DASD) [6] and our GSSVM over three events in TRECVID2010 MED dataset are reported.

6.2.2 Comparison with Fisher kernel

Table 2 compares the RBF-kernel GSSVM with a linear-kernel GSSVM and the improved Fisher kernel (IFK) [46]. For IFK, L_2 normalization and power normalization are applied to a Fisher vector. The parameter α of the power normalization is set to 0.5 as reported in [46]. As can be seen, the GSSVM with an RBF kernel outperformed the others. However, for some applications such as real-time event detection, the linear kernel is more reasonable than the RBF kernel in terms of scalability since it only requires a calculation of inner products to obtain a detection score.

6.2.3 Performance of camera motion cancelled feature

Table 3 shows the mean MinNDCs of each of the three types of motion features and combination of them: STIP, CC-DSTIP, and DSTIP in MED11 dataset. DSTIP is the dense STIP features without our camera motion cancellation.

As the single features, CC-DSTIP outperformed DSTIP in six events. The differences between CC-DSTIP and DSTIP are expected to be the effectiveness of camera motion cancellation. The major causes to make these differences are (1) the performance of camera motion cancellation and (2) the sizes of objects or motions in video clips. Events which have large improvement by camera motion cancellation are *Changing a vehicle tire* and *Repairing an appliance*. In these events, the foreground objects are often in the center part in a video clip and accordingly, the performance of camera motion cancellation is higher than other events. In *Parkour* and *Flash mob gathering*, dynamic motions such as *jumping* or *dancing* are included. On the other hand, in *Parade*, many people walking to one direction are often in video clips. In such a case, it is difficult to estimate camera motions since the peripheral regions in a video clip often include foreground motions. In *Birthday party*, *Grooming an animal*, or *Making a sandwich*, since foreground objects

Table 2 Comparison of the RBF-kernel GSSVM, the linear-kernel GSSVM, and the improved Fisher kernel [46]

Method	Mean MinNDC
RBF-kernel GSSVM	0.680
Linear-kernel GSSVM	0.693
Improved Fisher kernel	0.703

Minimum normalized detection cost (MinNDC) on TRECVID2010 MED for each kernel method is reported. STIP is used for low-level features.

Table 3 Comparison of STIP, DSTID, and CC-DSTIP

Event	STIP	DSTIP	CC-DSTIP	STIP + DSTIP	STIP + CC-DSTIP	DSTIP + CC-DSTIP	STIP +DSTIP +CC-DSTIP
Birthday party	*0.829*	0.853	0.873 (+0.020)	0.839	*0.807*	0.831	*0.820*
Changing a vehicle tire	0.834	0.844	*0.777* (−0.067)	0.810	*0.760*	0.783	*0.762*
Flash mob gathering	*0.441*	0.500	0.455 (−0.045)	0.436	*0.388*	0.435	*0.387*
Getting a vehicle unstuck	0.697	0.683	*0.664* (−0.019)	0.664	*0.619*	0.638	*0.608*
Grooming an animal	*0.785*	0.842	0.859 (+0.017)	0.778	*0.773*	0.782	*0.754*
Making a sandwich	*0.823*	0.876	0.873 (+0.003)	0.811	*0.785*	0.836	*0.767*
Parade	0.610	*0.597*	0.673 (+0.076)	0.580	*0.576*	0.594	*0.556*
Parkour	*0.425*	0.497	0.449 (−0.048)	0.444	*0.413*	0.468	*0.446*
Repairing an appliance	0.568	0.594	*0.533* (−0.061)	0.549	*0.537*	0.565	*0.541*
Working on a sewing project	*0.760*	0.777	0.781 (+0.004)	0.722	*0.688*	0.762	*0.722*
Mean	*0.677*	0.706	0.694 (−0.012)	0.663	*0.635*	0.669	*0.636*

Minimum NDCs of (1) three motion features, (2) combination of two of them, and (3) combination of three of them for each TRECVID2011 MED events are reported. The italic figures indicate the feature with the best performance in each category. Figures inside the parenthesis indicate the difference between CC-DSTIP and DSTIP.

such as *cake*, *animal*, and *sandwich* in video clips are often small in video clips compared with those of other events, the effectiveness of camera motion cancellation is not high.

Comparing STIP with DSTIP, STIP outperformed DSTIP in eight events. While recent studies have reported that dense sampling often outperforms keypoint-based sparse sampling (e.g., dense HOG in [33]), this shows that DSTIP does not always perform better than STIP as also shown in [25] on the KTH actions dataset. This is because the histogram of optical flow (HOF) in STIP extracted from flat regions is often noisy.

In the combination of two features, STIP and CC-DSTIP outperformed other combinations in all the events. Further, it outperformed the combination of all the three features in five events. It means CC-DSTIP provided us complementary information to STIP while the information of DSTIP was slightly complementary to STIP. This difference should be the effectiveness of camera motion cancellation.

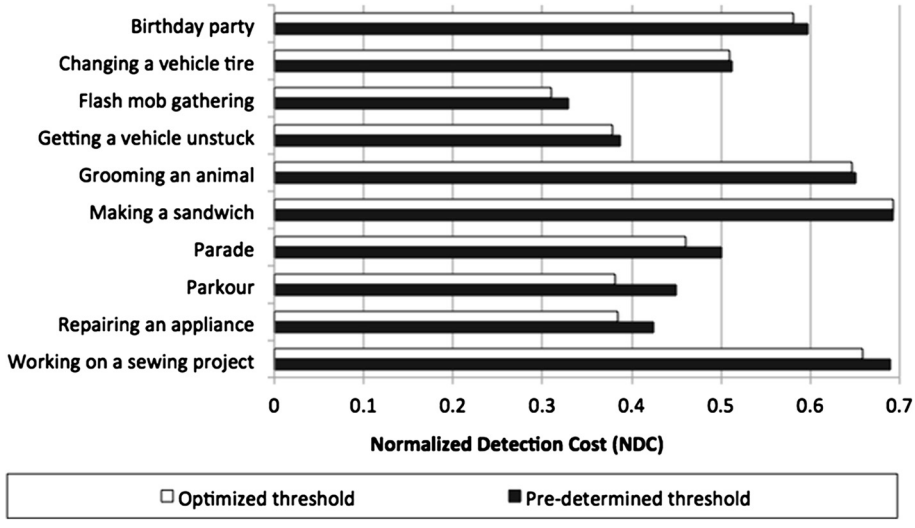

Figure 5 Normalized detection cost (NDC) for each event in TRECVID2011 MED. White bars: minimum NDC obtained using the most optimal thresholds for each event; black bars: NDC obtained using our thresholds.

6.2.4 Error ratios for TRECVID2011 MED events

Figure 5 shows NDC for each of the ten events in MED11. Looking at the results for each event, *Flash mob gathering* was the best event and *Making a sandwich* was the worst one. It is because *Making a sandwich* was difficult to distinguish from similar types of undefined events such as *Making a pizza* and *Making a cookie*, since their difference is only whether *a sandwich* appears or not in video clips. Figure 6 shows the detection result of *Making a sandwich*. Some of the video clips incorrectly detected have a kitchen but do not have a sandwich. For such events, semantic event model may be a desirable way to detect.

The comparison of NDC and MinNDC gives us how close the detection threshold is to the optimized one under the NDC criteria. The mean NDC with predetermined thresholds was 5.6% greater than that with the optimized threshold (MinNDC).

6.2.5 Our systems in TRECVID2011 MED

Figure 7 shows the comparison of our results to the median, average, and minimum (best) in original TRECVID2011 MED task. In total, 60 runs from 19 teams were submitted in this task. Our results consist of four types of fusions: visual features (Visual), audio feature (Audio), motion features (Motion), and all the features (All). CC-DSTIP was included in motion features. The performance of all the features performed significantly better than the median and average in all the events. In *Changing a vehicle tire* and *Getting a vehicle unstuck*, visual features showed the large gain compared to audio and motion features. It is because visual features captured the concept *car*, which characterizes these events. Motion features were effective particularly in *Parkour*, which often include *jumping*. Audio features performed better than the motion features in *Birthday party*, which often include *singing*. While each of visual, audio, and motion features showed some effectiveness by themselves in such events, they showed improvement when added with other types of features.

7 Conclusions

In the general event detection from consumer-generated videos, we do not have enough training data due to a

Figure 6 The top 25 clips in detection of *Making a sandwich*. The video clip on the first row and the first column has the highest detection score, the video clip on the first row and the second column has the second highest score, and the video clip on the second row and first column has the sixth highest score. The video clips with black frame are incorrectly detected. Others are correctly detected.

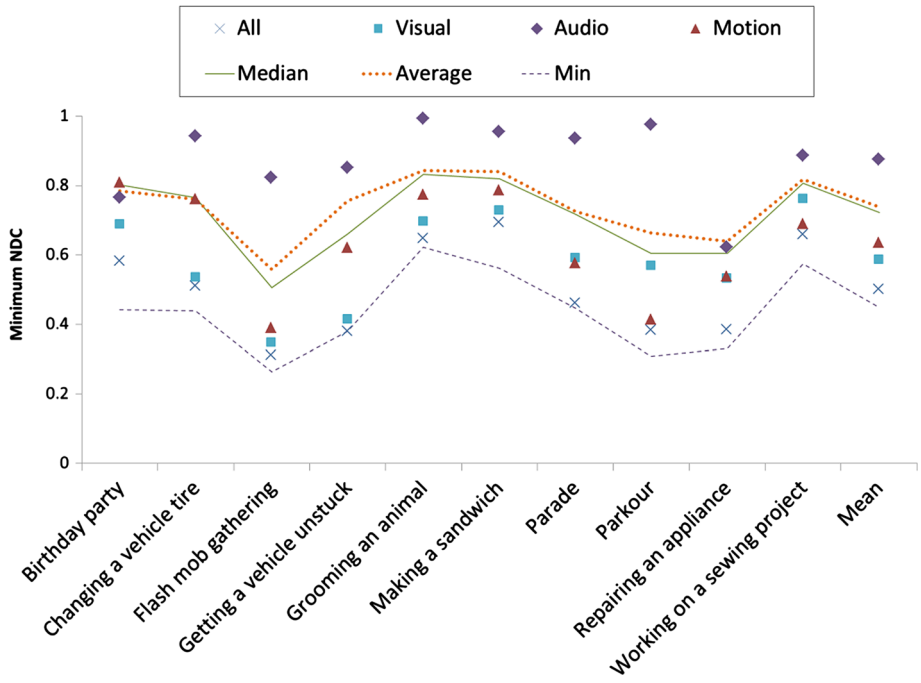

Figure 7 The comparison of minimum NDCs with TRECVID2011 MED runs. 'All' is the fusion of eight low-level features, 'Visual' is the fusion of five visual features, 'Audio' is MFCC features, and 'Motion' is the fusion of two motion features. 'Median,' 'Average,' and 'Min' is the median, average, and minimum (best) of MinNDCs over 60 runs submitted in TRECVID2011 MED task, respectively.

variety of video contents. To deal with this problem, we devised a general framework for multimedia event detection using GMM supervectors and SVMs (GSSVMs). Using a GMM, each clip is expected to be modelled precisely and GSSVM is expected to be robust against against the data insufficiency. Additionally, as one of the low-level features, we introduced camera motion cancelled features, which negate the affects of camera motions often included in consumer-generated videos. We combined GSSVM methods and camera motion cancelled features with seven types of existing complementary low-level features.

The GSSVM method performed better (mean MinNDC 0.510) than the previous studies using bag-of-words, domain-adaptive semantic diffusion, and earth mover's distance (mean MinNDC 0.565) in TRECVID2010 MED dataset. Further, our camera motion cancelled dense STIP features outperformed dense STIP features without the cancellation. As we assumed that camera motions were often useful for event detection, the combination of camera motion cancelled features and non-cancelled features were effective. The combination of multiple features, GMM supervectors, and SVMs showed effectiveness in the comparison of other methods in TRECVID2011 MED task.

For the future work, we will compare camera motion cancelled features with other motion features and camera motion cancellation methods. Other work includes detecting where an event occurs within a clip. Speeding up the GMM estimation and SVM training are also important work for large-scale event detection.

Endnote

[a] The work presented in this paper extends our prior approach published in [18].

Competing interests
The authors declare that they have no competing interests.

Received: 16 January 2013 Accepted: 30 July 2013
Published: 2 September 2013

References

1. J Assfalg, M Bertini, A Del Bimbo, W Nunziati, P Pala, Soccer highlights detection and recognition using HMMs, in *IEEE International Conference on Multimedia and Expo, 2002* (IEEE, Lausanne, 26–29 August 2002), pp. 825–828
2. Y Li, S Narayanan, C Kuo, Content-based movie analysis and indexing based on AudioVisual cues. Circuits Syst. Video Tech. IEEE Trans. **14**(8), 1073–1085 (2004)
3. A Adam, E Rivlin, I Shimshoni, D Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors. Pattern Anal. Mach. Intell. IEEE Trans. **30**(3), 555–560 (2008)
4. A Smeaton, P Over, W Kraaij, Evaluation campaigns and TRECVid, in *Proceedings of ACM International Workshop on Multimedia Information Retrieval, 2006* (ACM, Santa Barbara, CA, 26–27 October 2006), pp. 321–330
5. The National Institute of Stantdards and Technology (NIST), TRECVid multimedia event detection evaluation track. http://www.nist.gov/itl/iad/mig/med.cfm 2009. Accessed 15 Jan 2013
6. Y Jiang, X Zeng, G Ye, S Bhattacharya, D Ellis, M Shah, S Chang, Columbia-UCF TRECVID2010 multimedia event detection: combining multiple

modalities, contextual concepts, and temporal matching, in *Proceedings of TRECVID 2010 workshop* (NIST, Gaithersburg, MD, November 2010)
7. M Hill, G Hua, A Natsev, J Smith, L Xie, B Huang, M Merler, H Ouyang, M Zhou, IBM research TRECVID-2010 video copy detection and multimedia event detection system, in *Proceedings of TRECVID 2010 workshop* (NIST, Gaithersburg, MD, November 2010)
8. L Bao, S Yu, Z Lan, A Overwijk, Q Jin, B Langner, M Garbus, S Burger, F Metze, A Hauptmann, Informedia@ TRECVID 2011, in *Proceedings of TRECVID 2011 workshop* (NIST, Gaithersburg, MD, December 2011)
9. P Natarajan, S Wu, S Vitaladevuni, X Zhuang, S Tsakalidis, U Park, R Prasad, Multimodal feature fusion for robust event detection in web videos, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012* (IEEE, Providence, RI, 16–21 June 2012), pp. 1298–1305
10. G Csurka, CR Dance, L Fan, J Willamowski, C Bray, Visual categorization with bags of keypoints, in *Proceedings of IEEE European Conference on Computer Vision, 2004* (IEEE, Prague, 11–14 May 2004), pp. 59–74
11. J Yang, YG Jiang, AG Hauptmann, CW Ngo, Evaluating bag-of-visual-words representations in scene classification, in *Proceedings of ACM Multimedia MIR Workshop, 2007* (ACM, Augsburg, 24–29 September 2007), pp. 197–206
12. Z Ma, A Hauptmann, Y Yang, N Sebe, Classifier-specific intermediate representation for multimedia tasks, in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (ACM, Hong Kong, 05–08 June 2012), p. 50
13. Y Jiang, J Wang, S Chang, C Ngo, Domain adaptive semantic diffusion for large scale context-based video annotation, in *Proceedings of IEEE International Conference on Computer Vision, 2009* (IEEE, Kyoto, 27 September–04 October2009), pp. 1420–1427
14. N Inoue, K Shinoda, A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors. Multimedia, IEEE Trans. **14**(4), 1196–1205 (2012)
15. N Inoue, K Shinoda, A fast MAP adaptation technique for GMM-supervector-based video semantic indexing systems, in *Proceedings of ACM Multimedia, 2011* (ACM, Scottsdale, AZ, 28 November–01 December 2011), pp. 1357–1360
16. Y Liu, F Perronnin, A similarity measure between unordered vector sets with application to image categorization, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008* (IEEE, Anchorage, AL, 24–26 June 2008)
17. X Zhou, K Yu, T Zhang, T Huang, Image classification using super-vector coding of local image descriptors, in *Proceedings of IEEE European Conference on Computer Vision, 2010* (IEEE, Heraklion, 5–11 September 2010), pp. 141–154
18. Y Kamishima, N Inoue, K Shinoda, S Sato, Multimedia event detection using GMM supervectors and SVMS, in *Proceedings of IEEE International Conference on Image Processing, 2012* (IEEE, Orlando, FL, 30 September–03 October 2012), pp. 3089–3092
19. D Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
20. X Zhou, X Zhuang, S Yan, S Chang, M Hasegawa-Johnson, T Huang, SIFT-bag kernel for video event analysis, in *Proceedings of ACM Multimedia, 2008* (ACM, Vancouver, 27–31 October 2008), pp. 229–238
21. E Nowak, F Jurie, B Triggs, Sampling strategies for bag-of-features image classification, in *Proceedings of IEEE European Conference on Computer Vision, 2006* (IEEE, Austria, 7–13 May 2006), pp. 490–503
22. H Bay, A Ess, T Tuytelaars, L Van Gool, SURF: speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
23. N Inoue, S Tatsuhiko, K Shinoda, S Furui, High-level feature extraction using SIFT GMMs and audio models, in *Proceedings of International Conference on Pattern Recognition, 2010* (IAPR, Istanbul, 23–26 August 2010), pp. 3220–3223
24. I Laptev, On space-time interest points. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
25. H Wang, M Ullah, A Klaser, I Laptev, C Schmid, Evaluation of local spatio-temporal features for action recognition, in *Proceedings of British Machine Vision Conference, 2009* (BMVA, London, 7–10 September 2009)
26. A Kläser, M Marszalek, C Schmid, A spatio-temporal descriptor based on 3Dgradients, in *Proceedings of British Machine Vision Conference, 2008* (BMVA, Leeds, September 2008)
27. M Chen, A Hauptmann, Mosift: recognizing human actions in surveillance videos, in *CMU-CS-09-161* (Carnegie Mellon University, 2009)
28. S Wu, O Oreifej, M Shah, Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories, in *Proceedings of IEEE International Conference on Computer Vision, 2011* (IEEE, Barcelona, 6–13 November 2011), pp. 1419–1426
29. H Wang, A Kläser, C Schmid, CL Liu, Action recognition by dense trajectories, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011* (IEEE, Colorado Springs, CO, 20–25 June 2011), pp. 3169–3176
30. K Mikolajczyk, H Uemura, Action recognition with appearance - motion features and fast search trees. Comput. Vis. Image Underst. **115**(3), 426–438 (2011)
31. N Ikizler-Cinbis, S Sclaroff, Object, scene and actions: combining multiple features for human action recognition, in *Proceedings of IEEE European Conference on Computer Vision, 2010* (IEEE, Heraklion, 5–-11 September 2010), pp. 494–507
32. K Li, S Oh, A Perera, Y Fu, A videography analysis framework for video retrieval and summarization, in *BMVC* (BMVA, 2012)
33. N Inoue, T Wada, Y Kamishima, K Shinoda, S Sato, TokyoTech+Canon at TRECVID 2011, in *Proceedings of TRECVID 2011 workshop* (NIST, Gaithersburg, MD, December 2011)
34. K Mikolajczyk, C Schmid, Scale & affine invariant interest point detectors. Int. J. Comput. Vis. **60**, 63–86 (2004)
35. N Dalal, B Triggs, C Schmid, Human detection using oriented histograms of flow and appearance, in *Proceedings of IEEE European Conference on Computer Vision, 2006* (IEEE, Austria, 7–13 May 2006), pp. 428–441
36. KEA van de Sande, T Gevers, CGM Snoek, Evaluating color descriptors for object and scene recognition. Pattern Anal. Mach. Int. IEEE Trans. **32**(9), 1582–1596 (2010)
37. S Lazebnik, C Schmid, J Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories , in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006* (IEEE, New York, NY, 17–22 June 2006), pp. 2169–2178
38. W Campbell, D Sturim, D Reynolds, A Solomonoff, SVM based speaker verification using a GMM supervector kernel and nap variability compensation, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006* (IEEE, Graz, 14–19 May 2006), pp. 97–100
39. J Gauvain, CH Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. Speech Audio Proc. IEEE Trans. **2**(2), 291–298 (1994)
40. A Vedaldi, SIFT++. http://www.vlfeat.org/∼vedaldi/code/siftpp.html 2006. Accessed 15 Jan 2013
41. SJ Young, G Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, D Valtchev, PC Woodland, The hkt book. http://htk.eng.cam.ac.uk 2006. Accessed 15 Jan 2013
42. CC Chang, CJ Lin, Libsvm: A library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/ 2001. Accessed 15 Jan 2013
43. DG Lowe, Object recognition from local scale-invariant features , in *Proceedings of IEEE International Conference on Computer Vision, 1999* (IEEE, Corfu, 20–25 September 1999), pp. 1150–1157
44. Y Rubner, C Tomasi, L Guibas, The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
45. A Smeaton, P Over, W Kraaij, High-level feature detection from video in TRECVid: a 5-year retrospective of achievements, in *Multimedia Content Analysis* (Springer US, Norwell, 2009), pp. 151–174
46. F Perronnin, J Sánchez, T Mensink, Improving the fisher kernel for large-scale image classification, in *Proceedings of IEEE European Conference on Computer Vision, 2010* (IEEE, Heraklion, 5–11 September 2010), pp. 143–156

doi:10.1186/1687-5281-2013-51
Cite this article as: Kamishima *et al.*: **Event detection in consumer videos using GMM supervectors and SVMs.** *EURASIP Journal on Image and Video Processing* 2013 **2013**:51.