Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment

Dong Xu, *Member*, *IEEE*, and Shih-Fu Chang, *Fellow*, *IEEE*

Abstract—In this work, we systematically study the problem of event recognition in unconstrained news video sequences. We adopt the discriminative kernel-based method for which video clip similarity plays an important role. First, we represent a video clip as a bag of orderless descriptors extracted from all of the constituent frames and apply the earth mover's distance (EMD) to integrate similarities among frames from two clips. Observing that a video clip is usually comprised of multiple subclips corresponding to event evolution over time, we further build a multilevel temporal pyramid. At each pyramid level, we integrate the information from different subclips with Integer-value-constrained EMD to explicitly align the subclips. By fusing the information from the different pyramid levels, we develop **T**emporally **A**ligned **P**yramid **M**atching (TAPM) for measuring video similarity. We conduct comprehensive experiments on the TRECVID 2005 corpus, which contains more than 6,800 clips. Our experiments demonstrate that 1) the TAPM multilevel method clearly outperforms single-level EMD (SLEMD) and 2) SLEMD outperforms keyframe and multiframe-based detection methods by a large margin. In addition, we conduct in-depth investigation of various aspects of the proposed techniques such as weight selection in SLEMD, sensitivity to temporal clustering, the effect of temporal alignment, and possible approaches for speedup. Extensive analysis of the results also reveals intuitive interpretation of video event recognition through video subclip alignment at different levels.

Index Terms—Event recognition, news video, concept ontology, Temporally Aligned Pyramid Matching, video indexing, earth mover's distance.

1 Introduction

Event recognition from visual cues is a challenging task because of complex motion, cluttered backgrounds, occlusions, and geometric and photometric variances of objects. Previous work on video event recognition can be roughly classified as either activity recognition or abnormal event recognition.

For model-based abnormal event recognition, Zhang et al. [1] propose a semisupervised adapted Hidden Markov Model (HMM) framework in which usual event models are first learned from a large amount of training data and unusual event models are learned by Bayesian adaptation. In model-based approaches to activity recognition, frequently used models include HMM [2], coupled HMM [3], and Dynamic Bayesian Network [4]. The work in [5] modeled each activity with a nominal activity trajectory and one function space for time warping. To model the relationship between different parts or regions, object tracking is usually performed before model learning [2], [3], [6]. Additionally, these techniques heavily rely on the choice of good models, which in turn requires sufficient training data to learn the model parameters.

- *D. Xu is with the School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Blk N4, Singapore, 639798. E-mail: dongxu@ntu.edu.sg.*
- *S.-F. Chang is with the Department of Electrical Engineering, Columbia University, 500 W. 120th St. Rm 1312, New York, NY 10027. E-mail: sfchang@ee.columbia.edu.*

Manuscript received 4 Sept. 2007; revised 4 Feb. 2008; accepted 24 Mar. 2008; published online 20 May 2008.
Recommended for acceptance by J.Z. Wang, D. Geman, J. Luo, and R.M. Gray. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2007-09-0550.
Digital Object Identifier no. 10.1109/TPAMI.2008.129.

Appearance-based techniques extract spatiotemporal features in the volumetric regions, which can be densely sampled or detected by salient region detection algorithms. For abnormal event recognition, Boiman and Irani [7] proposed extracting an ensemble of densely sampled local video patches to localize irregular behaviors in videos. For activity recognition, Ke et al. [8] applied boosting to choose volumetric features based on optical flow representations. Efros et al. [9] also used optical flow measurements in spatiotemporal volumetric regions. Other researchers extracted volumetric features from regions with significant local variations in both spatial and temporal dimensions [10], [11], [12], [13]. Laptev and Lindeberg [11] detected salient regions by extending the idea of Harris interest point operators [14] and Dollar et al. [10] applied separable linear filters for the same objective. Learning techniques such as support vector machine (SVM) [13] and probabilistic latent semantic analysis (pLSA) [12] are also combined with the above representation [10], [11] to further improve performance. The performance of appearance-based techniques usually depends on reliable extraction of spatial-temporal features and/or salient regions, which are often based on optical flow or intensity gradients in the temporal dimension. This makes the approach sensitive to motion, e.g., the detected interest regions are often associated with high-motion regions [8].

The extensive works mentioned above have demonstrated promise for event detection in domains such as surveillance or constrained live video. However, generalization to less constrained domains like broadcast news has not been demonstrated. Broadcast news videos contain rich information about objects, people, activities, and events. Recently, due to the emerging applications of open source intelligence and online video search, the research community has shown

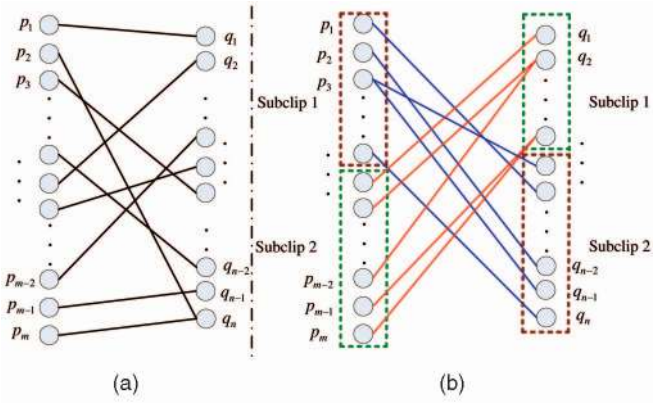

Fig. 1. Illustration of temporal matching in (a) SLEMD and (b) multilevel TAPM.

increasing interest in event recognition in broadcast news videos. For example, analysts or ordinary users may be interested in specific events (e.g., a visit of a government leader and a world trade summit) or general event classes. To respond to such interests, Large-Scale Concept Ontology for Multimedia (LSCOM) ontology [15], [16] has defined 56 event/activity concepts, covering a broad range of events such as car crash, demonstration, riot, running, people marching, shooting, walking, and so on. These events were selected through a triage process based on input from a large group of participants, including video analysts, knowledge representation experts, and video analysis researchers. Manual annotation of such event concepts has been completed for a large data set in TRECVID 2005 [15].

Compared with prior video corpora used in abnormal event recognition and activity recognition, news videos are more diverse and challenging due to the large variations of scenes and activities. Events in news video may involve small objects located in arbitrary locations in the image under large camera motions. Therefore, it is difficult to reliably track moving objects in news video, detect the salient spatiotemporal interest regions, and robustly extract the spatial-temporal features.

To address the challenges of news video, Ebadollahi et al. [17] proposed treating each frame in the video clip as an observation and applying an HMM to model the temporal patterns of event evolution. Such approaches are distinct from most prior event recognition techniques since they circumvent the need for object tracking. In contrast, holistic features are used to represent each frame and the focus is on the modeling of temporal characteristics of events. However, results shown in [17] did not confirm clear performance improvements over a simplistic detection method using static information in keyframes only. This is perhaps due to the lack of a large training set required to learn the model parameters.

In this work, we also adopt holistic representations for image frames without object tracking or spatiotemporal interest region detection. We propose using midlevel concept score (CS) features to abstract visual content in unconstrained broadcast news video. Each CS is generated by a semantic concept classifier, which can represent the

Fig. 2. Illustration of our midlevel concept scores. Scores shown in graph bars are actual values generated by five concept detectors: anchor, snow, soccer, building, and outdoor.

semantic meaning of the image to some extent, as shown in Fig. 2 and discussed in detail in Section 2. We also propose a nonparametric approach in order to circumvent the problems of insufficient training data for model learning and to simplify the training process. We investigate how to efficiently utilize the information from multiple frames, as well as the temporal information within each video clip. Inspired by the recent success of the bag-of-words model in object recognition [18], [19], [20], we first represent one video clip as a bag of orderless features, extracted from all of the frames. To handle the temporal variations in different video clips, we apply the earth mover's distance (EMD) [21], referred to as single-level EMD (SLEMD) in this work, to measure video clip similarity and combine it with SVM kernel classifiers for event detection. The idea behind EMD has been used for decades under a variety of names in probability, statistics, ergodic theory, information theory, and so on, see [22], [23], [24] for more details. SLEMD computes the optimal flows between two sets of frames, yielding the optimal match between two frame sets, as shown in Fig. 1a.

We also observe that one video clip is usually comprised of several subclips which correspond to multiple stages of event evolution. For example, Figs. 3a and 3f show two news video clips in the *Riot* class, both of which consist of distinct stages of fire, smoke, and/or different backgrounds. Given such multistage structures, it makes sense to extend the SLEMD mentioned above to multiple scales. Specifically, we propose a Temporally Aligned Pyramid Matching (TAPM) framework in which temporal matching is performed in a multiresolution fashion. As shown in Fig. 1b,

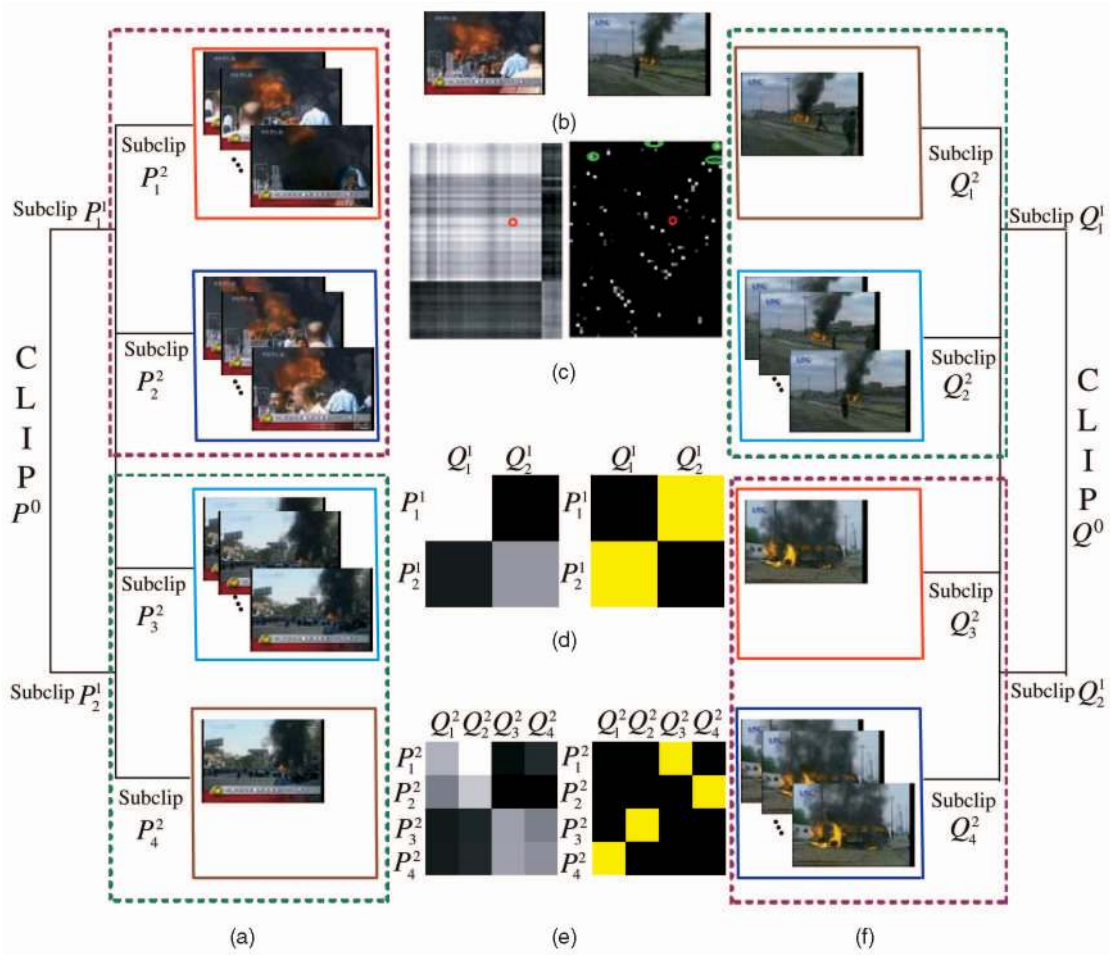

Fig. 3. Conceptual illustration for Temporally Aigned Pyramid Matching. For better viewing, please see the color pdf file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.129. (a) and (f) Frames from two videos, P and Q, from the *Riot* class are segmented into multiple subclips. (b) One keyframe for each clip. (c) Ground distance matrix and continuous-value flow matrix from the SLEMD alignment process. The red circles indicate the locations of the keyframes in the corresponding video clips. The green circles highlight the problem that one frame may be matched to multiple frames that are far apart. (d) and (e) The EMD distance matrices between the subclips and their corresponding integer-value flow matrices are shown at level-1 and level-2. In contrast to (c), the first several frames in P can only be matched to frames within a subclip of the same border color in Q; thus, temporal proximity is preserved.

frames in a subclip (conceptually corresponding to an event stage) are matched to frames within *one* subclip in the other video, rather than spread over *multiple* subclips in Fig. 1a. Such constraints may seem to be overrestrictive, but it explicitly utilizes the temporal information and is experimentally demonstrated to be effective for improving detection accuracy of several events in Section 5.3. Furthermore, there is no prior knowledge about the number of stages in an event and videos of the same event may include a subset of stages only. To address this problem, we propose fusing the matching results from multiple temporal scales in a way similar to that used in Spatial Pyramid Matching (SPM) [18] and Pyramid Match Kernel (PMK) [25] for object recognition. However, it is worthwhile to point out that SPM used fixed block-to-block matching in the *spatial* domain for scene classification based on the observation that images from the same scene have similar spatial configurations. But, in TAPM, the subclips across different *temporal* locations may be matched, as shown in Fig. 1b, which will be explained in detail in Section 4.2.

We conduct comprehensive experiments on the large TRECVID 2005 database and the experiments demonstrate that 1) the multilevel matching method TAPM clearly outperforms SLEMD and 2) SLEMD outperforms basic detection methods that use one keyframe or multiple keyframes. To the best of our knowledge, this work and our initial conference version [26] are the first to systematically study the problem of visual event recognition in less constrained video domains such as broadcast news without any specific parametric model or object tracking.

2 Scene-Level Concept Score Feature

In this work, we adopt holistic features to represent content in constituent image frames and a multilevel temporal alignment framework to match temporal characteristics of various events. We consider three low-level global features—Grid Color Moment, Gabor Texture, and Edge Direction Histogram—because of their consistent, good performance reported in TRECVID [27], [28]. For the Grid Color Moment feature, we extract the first three moments of three channels in

the CIE Luv color space from each of the 5×5 fixed grid partitions and aggregate the features into a single 225-dimensional feature vector. For the Gabor Texture feature, we take four scales and six orientations of the Gabor transformation and compute their means and standard deviations, giving a dimension of 48. The Edge Direction Histogram feature includes 73 dimensions with 72 bins corresponding to edge directions quantized at five degrees and one bin for nonedge pixels. For more details about the low-level global features, please refer to [27], [28]. Note that similar color, texture, and edge features were frequently used in TRECVID by other research groups [29], [30], [31], [32].

We also use a large number of midlevel semantic CSs to abstract visual information. Such a midlevel CS feature has shown promise for abstracting visual content and significantly improving the performance of semantic video retrieval [33]. To detect the semantic concepts in general broadcast news videos, we have developed 374 concept detectors, Columbia374-baseline [33], [34], to detect a diverse set of semantic concepts. These concepts are carefully chosen from the LSCOM ontology [15] and are related to important categories of information in open source news videos, such as events, objects, locations, people, and programs. The complete listing of the LSCOM ontology, the 374 semantic concepts, and their annotations over the entire set of video shots from the TRECVID 2005 data set (61,901 shots from 137 video programs) can be found in [15]. In Columbia374-baseline [33], [34], the training data are from the keyframes of 90 video programs in the TRECVID 2005 data set. Based on the above three low-level features, three independent SVMs were trained and then fused to produce the scores for each concept. Each image is then mapped to a semantic vector space in which each element presents a confidence score (or response) produced by a semantic concept classifier. For more details about Columbia374-baseline, please refer to [33], [34].

Fig. 2 shows some example images and their responses from five concept detectors: anchor, snow, soccer, building, and outdoor. We observe that our midlevel CS feature can adequately characterize the semantic meaning of an image to some extent. For example, for the images in the third row, the responses from concept detectors building and outdoor are strong, but the responses from the other three detectors are weak, which reasonably represents the semantic meaning of these images. In this work, we leverage the potential of large-scale concept detectors and devise effective mechanisms of event recognition based on such a concept representation framework. Columbia374 baseline models have been released to the community [34] with the goal of fostering innovation in concept detection and enabling the exploration of the use of a large set of concept detectors for other tasks (e.g., video search and event recognition).

While it is possible to use other local features, such as term frequency-inverse document frequency (tf-idf) features [19] based on SIFT local descriptors [35], we use the above global features and scene-level CSs because 1) they can be efficiently extracted over the large video corpus while avoiding the difficult task of object detection and tracking, 2) they have been shown to be effective for detecting several concepts in previous TRECVID experiments [27], [28], [29], [30], [31], [32], and 3) they are suitable for capturing the characteristics of scenes in some events such as *Riot*, *Car Crash*, *Exiting Car*, and so on. Examples of the dominant scenes for the *Riot* class are shown in Fig. 3 and the experimental results shown in Section 5 also confirm the potential of the representations.

3 Single-Level Earth Mover's Distance in the Temporal Domain

We represent a video clip as a bag of orderless descriptors (e.g., midlevel concept scoress) extracted from all of the constituent frames and we aim to compute the distance between two sets of points with unequal length. In the literature, several algorithms (e.g., the Kullback-Leibler Divergence-based Kernel [36], Pyramid Match Kernel [25], and EMD [21]) have been proposed. In this work, we apply the EMD [21] to integrate similarities among frames from two clips because of its promising performance in several different applications such as content-based image retrieval [21], [37], texture classification, and general object recognition [20]. It is worthwhile to highlight that our framework is a general framework for event recognition in unconstrained news video and other techniques, such as [25], [36], may be combined with EMD to further improve the performance.

In this section, we develop an SLEMD to efficiently utilize the information from multiple frames of a video clip for event recognition. We will extend this method to multiple levels in Section 4. One video clip P can be represented as a signature: $P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$, where m is the total number of frames, p_i is the feature extracted from the ith frame, and w_{p_i} is the weight of the ith frame. The weight w_{p_i} is used as the total supply of suppliers or the total capacity of consumers in the EMD method, with the default value of $1/m$. p_i can be any feature, such as Grid Color Moment [27] or CS feature [33]. We also represent another video clip Q as a signature: $Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$, where n is the total number of frames and q_i and $w_{q_i} = 1/n$ are defined similarly. We note the same setting was also used in [20], [37] because of its simplification and efficiency. In news video, it is a challenging task to handle temporal shifts and duration variations. With this setting, we observe that one frame at the beginning of P may be matched to one frame at the end of Q, which indicates that EMD with default normalized weights can efficiently handle temporal shifts. Also, as shown in Fig. 3c, flow values computed with (2) can be continuous values; therefore, one frame of P may be matched to multiple frames of Q, which indicates that it can handle duration variations to some extent. Moreover, in Section 5.2, we demonstrate through experiments that EMD with the normalized value works better than other possible weights, e.g., unit weight. The EMD between P and Q can be computed by

$$D(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \widehat{f}_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} \widehat{f}_{ij}}, \quad (1)$$

where d_{ij} is the ground distance between p_i and q_j (we use euclidian distance as the ground distance in this work

because of its simplicity and prior success reported in [20], [37]) and $\widehat{f}_{ij}$ is the optimal flow that can be determined by solving the following Linear Programming problem with the standard Simplex method [21].

$$\widehat{f}_{ij} = arg\min_{f_{ij}} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}$$
$$s.t. \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right); \quad f_{ij} \geq 0;$$
$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, 1 \leq i \leq m; \sum_{i=1}^{m} f_{ij} \leq w_{q_j}, 1 \leq j \leq n. \tag{2}$$

$\widehat{f}_{ij}$ can be interpreted as the optimal match among frames from two video clips, as shown in Fig. 1a. Since euclidean distance is a metric and the total weight of each clip is constrained to be 1, the EMD distance is therefore a true distance because nonnegativity, symmetry, and the triangle inequality all hold in this case [21]. Suppose the total number of frames in two video clips are the same, i.e., $m = n$, then the complexity of EMD is $O(m^3 log(m))$ [21].

Figs. 3a and 3f show all of the frames of the two video clips P and Q from the *Riot* class. Fig. 3b shows two representative keyframes provided by the TRECVID data set. The ground distance matrix and the flow matrix between the frames of P and Q are shown in Fig. 3c, in which the brighter pixels indicate that the values at that position are larger. In Fig. 3c, we also use the red circles to indicate the positions of two keyframes. From this, we can see that the flow (calculated from the EMD process) between these two keyframes is very small, confirming that a keyframe-based representation is not sufficient for capturing characteristics over multiple frames of the same event class. We also use four green circles in Fig. 3c to highlight that the neighboring frames from the first stage of P are matched to distant frames scattered among several different stages of Q. In other words, temporal proximity and order among frames are not preserved in this SLEMD method. In Section 4, we will propose a multiresolution framework to partially preserve the proximity relations.

For classification, we use an SVM because of its good performance [18], [20], [25], [37]. For a two-class (e.g., *Riot* versus *non-Riot*) case, we use the Gaussian function to incorporate EMD distance from every pair of training video clips into the SVM classification framework:

$$K(P,Q) = exp\left(-\frac{1}{A}D(P,Q)\right). \tag{3}$$

In the training stage, we set hyperparameter A to κA_0, where the normalization factor A_0 is the mean of the EMD distances between all training video clips and the optimal scaling factor κ is empirically decided through cross validation. While no proof exists for the positive definiteness of the EMD-kernel, in our experiments, this kernel has always yielded positive definite Gram matrices. Furthermore, as shown in [20], [37], EMD-kernel works well in content-based image retrieval and object recognition. In the testing stage, the decision function for a test sample P has the following form:

$$g(P) = \sum_{t} \alpha_t y_t K(P, Q_t) - b, \tag{4}$$

where $K(P, Q_t)$ is the value of a kernel function for the training sample Q_t and the test sample P, y_t is the class label of Q_t ($+1$ or -1), α_t is the learned weight of the training sample Q_t, and b is the threshold parameter. The training samples with weight $\alpha_t > 0$ are called support vectors. The support vectors and their corresponding weights are learned using the standard quadratic programming optimization process or other variations. In this project, we use tools from libsvm [38] in our implementations.

4 Temporally Aligned Pyramid Matching

We observe that one video clip is usually comprised of several subclips (subclip partition will be discussed in Section 4.1), which correspond to event evolution over multiple stages. For example, in Figs. 3a and 3f, we can observe that videos from the *Riot* class may consist of two stages, involving varying scenes of fire and smoke and different locations. Recent works such as SPM [18] and Pyramid Match Kernel [25] have demonstrated that better results may be obtained by fusing the information from multiple resolutions according to pyramid structures in the spatial domain and feature domain, respectively. Inspired by these works, we propose applying Temporally Constrained Hierarchical Agglomerative Clustering (T-HAC) to build a multilevel pyramid in the temporal domain. According to the multilevel pyramid structure, each video clip is divided into several subclips, which may correspond to distinct stages involved in the evolution process of the event. From now on, we denote the video clip at level-0 (i.e., the original video clip) as P^0 or P and the subclips at level-l as P_r^l, where $r = 1, \ldots, R = 2^l$, $l = 0, \ldots, L-1$ with L as the total number of levels. For example, as shown in Figs. 3a and 3f, four subclips at level-2 are denoted as P_1^2, P_2^2, P_3^2, and P_4^2, which are bounded by solid bounding boxes, and two subclips at level-1 are denoted as P_1^1 and P_2^1, which are bounded by a dashed bounding box.

We also observe that, in broadcast news videos, stages of two different clips of the same event in general may not follow a fixed temporal order. To address this problem, we integrate the information from different subclips with Integer-value-constrained EMD to explicitly align the orderless subclips (see Section 4.2). Finally, we fuse the information from different levels of the pyramid, which results in the TAPM.

4.1 Temporally Constrained Hierarchical Agglomerative Clustering

We use Hierarchical Agglomerative Clustering [39] to decompose a video into subclips. Clusters are constructed by iteratively combining existing clusters based on their distances. To incorporate temporal information, we propose using T-HAC, in which, at each step, we only merge neighboring clusters in the temporal dimension. More details are listed in Algorithm 1. Examples of the clustering results are also shown in Figs. 3a and 3f. We observe that T-HAC provides a reasonable pyramid structure in the temporal dimension.

Algorithm 1. T-HAC.

1. Each frame is initialized as a singleton cluster;
2. For any two neighboring clusters from the constituent frames, compute their distance in terms of the minimum temporal time difference criterion. Then, the two clusters with the minimal distance are merged together to form a new cluster and hence the total number of clusters is reduced by one;
3. This merging procedure continues until the prespecified number of clusters is reached.

Later, in Section 5.4, we will experimentally compare the results from T-HAC with a simple solution for dividing video clips into different stages, which uniformly partitions one clip into several subclips. We also analyze the sensitivity of our framework to the temporal clustering algorithm by adding random perturbations to the temporal boundary from T-HAC. We also note that the proposed framework is general and other clustering algorithms may be readily incorporated into our framework to replace the T-HAC clustering method described above. As for determination of the suitable number of clusters, we conjecture that there is no single best value and thus propose an effective method to fuse the subclip matching results from multiple resolutions (see Section 4.3) in a way similar to that adopted in SPM for object recognition [18].

4.2 Alignment of Different Subclips

After we build the pyramid structure in the temporal dimension for two video clips P and Q, we need to compute the distance at each level, i.e., $S^l(P,Q)$ at level l. First, we apply (1) to compute the EMD distances between subclips at level l, i.e., D^l. Figs. 3d and 3e (the left-hand-side matrices) show examples for P and Q at level-1 and level-2, respectively, in which again a higher intensity represents a higher value between the corresponding subclips. For example, at level-2, we compute a 4×4 EMD distance matrix with its elements denoted as D^2_{rc}, the EMD distance between P^2_r and Q^2_c, $r = 1, \ldots, 4$ and $c = 1, \ldots, 4$. If we assume a uniform partitioning of a video clip into subclips, the complexity involved in computing level-l distance is $O((2^{-l}m)^3 log(2^{-l}m))$, which is significantly lower than the SLEMD (i.e., $l = 0$). Therefore, fusing information from higher levels does not add significant computational cost to the overall detection method.

If we follow the same strategy as in SPM [18], which assumes that the corresponding subregions of any scene category are well aligned according to their position in the images, we can take the sum of the diagonal elements of D^l as the level-l distance. However, in Fig. 3d, we observe that the diagonal elements of D^l are very large because the first stage of P and the last stage of Q focus on the fire scene and the last stage of P and the first stage of Q mainly depict smoke and the nature scene. Thus, it is desirable to align the subclips shown in Figs. 3a and 3f inversely in the temporal domain. Similar observations can be found from the EMD distances at level-2 (a 4×4 matrix) in Fig. 3e. In Fig. 4, we also plot the data distribution in the three-dimensional space of all of the frames from two subclips of P and Q (both depicting *Riot*) at level-1 and from other negative video clips, e.g., videos from other events such as *People Walking*, *Ground Combat*, and so on. The CS feature is projected into 3D space by Principal Component Analysis (PCA). In Fig. 4, it is seen that video clips from event *Riot* may be comprised of two stages occupying consistent locations in the 3D space, except that the sequential order of the two stages may be switched, according to the data distribution in the 3D space.

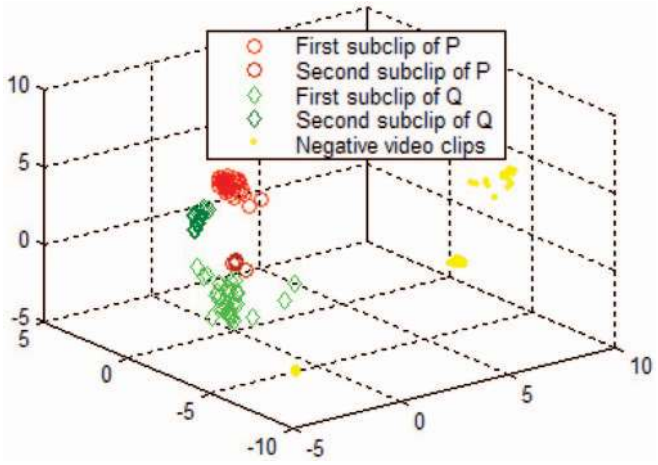

Fig. 4. Projection of image frames of two videos (P and Q) of the same event *Riot* into a 3D space via PCA. Note the projected points from two videos share similar locations centered at two distinct clusters, but the points from the first and second stages (subclips) are swapped between the two clusters. The yellow points are from the negative set (i.e., *non-Riot*) and thus do not share similar locations with those from the positive set. For better viewing, please see the color pdf file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.129.

To explicitly align the different subclips and utilize the temporal information, we constrain the linear programming problem in EMD to an integer solution (i.e., integer programming). Such an integer solution can be conveniently computed by using the standard Simplex method for Linear Programming, according to the following Theorem 1.

Theorem 1 [40]. *The Linear Programming problem,*

$$arg\min_{F_{rc}} \sum_{r=1}^{R}\sum_{c=1}^{C} F_{rc}D_{rc}, \quad s.t. \quad 0 \le F_{rc} \le 1, \quad \forall r, c; \\ \sum_{c} F_{rc} = 1, \quad \forall r; \quad \sum_{r} F_{rc} = 1, \quad \forall c; \quad \text{and} \quad R = C, \tag{5}$$

will always have an integer optimum solution when it is solved with the Simplex method.[1]

More details about the theorem can be found in [40]. We use the integer-value constraint to explicitly enforce the constraint that neighboring frames from a subclip are mapped to neighboring frames in a subclip in the other video. Without such an integer-value constraint, one subclip may be matched to several subclips, resulting in a problem that one frame is mapped to multiple distant

1. This problem can be also formulated as a minimum-weighted bipartite matching problem [41]. However, the main focus of this work is to propose a general framework for event recognition in unconstrained news video. To be consistent with our single-level EMD, we formulate it as an integer-value-constrained EMD and still use the Simplex method to solve it. Note tht the computation cost is low in this step because the total number of parameters is only 4 and 16 at level-1 and level-2, respectively.

frames (as exemplified by the green circles in Fig. 3c). Note that noninteger mappings are still permitted at level-0 so that soft temporal alignments of frames are still allowed within the video clip at level 0. Similarly to (1), the level-l distance $S^l(P,Q)$ between two clips P and Q can be computed from the integer-value flow matrix F^l and the distance matrix D^l by

$$S^l(P,Q) = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} F_{rc}^l D_{rc}^l}{\sum_{r=1}^{R}\sum_{c=1}^{C} F_{rc}^l}, \quad (6)$$

where $R = C = 2^l$ and F_{rc}^l are 0 or 1.

Again, we take the subclips at level-1 as an example. According to Theorem 1, we set $R = C = 2$ and obtain a 2×2 integer flow matrix F_{rc}^1, as shown in Fig. 3d (the right-hand-side matrix). From it, we can find that two subclips of P and Q are correctly aligned (i.e., inversely aligned). Similar observations at level-2 can be found in the right-hand-side matrix in Fig. 3e, where we set $R = C = 4$. In Figs. 3a and 3f, the aligned subclips are shown with the same border color.

Similarly to the kernel-based method described in Section 3, for each level l, we apply (3) to convert level-l distance $S^l(P,Q)$ to its kernel matrix. In the training stage, we train a level-l SVM model for each level l. In the testing stage, for each test sample P, according to (4), we can obtain the decision values $g^0(P)$, $g^1(P)$, $g^2(P)$, and so on, from the SVM models trained at different levels.

4.3 Fusion of Information from Different Levels

As shown in [18], the best results can be achieved when multiple resolutions are combined, even when some resolutions do not perform well independently. In this work, we directly fuse the decision values $g^l(P)$, $l = 0, \ldots, L-1$ from different level SVM models according to the method suggested in [33], [42]:

$$g^f(P) = \sum_{l=0}^{L-1} \frac{h_l}{1 + exp(-g^l(P))}, \quad (7)$$

where h_l is the weight for level-l.

In our experiments, we set $L = 3$. We tried two weighting schemes: 1) the weights suggested in [18] and [25], i.e., equal weights $h_0 = h_1 = 1$ to fuse the first two levels and diadic weights $h_0 = h_1 = 1$, $h_2 = 2$ to fuse all levels and 2) equal weights $h_0 = h_1 = h_2 = 1$ to fuse all three levels. In Section 5, we will present experimental results to demonstrate that the results with equal weights are comparable to or slightly better than the weights suggested in [18], [25] when fusing all three levels. More details about TAPM are listed in Algorithm 2.

Algorithm 2. TAPM.

1. Divide video clip P and Q into subclips P_r^l and Q_c^l, $l = 1, 2, \ldots, L-1$, with T-HAC;
2. Compute the distance $S^0(P,Q)$ with (1) and (2) and then SVM training and testing based on $S^0(P,Q)$;
3. For $l = 1, 2, \ldots, L-1$, Do
   1) Compute the EMD distances D_{rc}^l between subclips with (1) and (2);
   2) Compute the integer-value flow matrix F^l and the distance $S^l(P,Q)$ with (5) and (6);
   3) SVM training and testing based on $S^l(P,Q)$;
4. Fuse the information from multiple resolutions with (7).

5 Experiments

We conduct comprehensive experiments over the large TRECVID 2005 video corpus to compare 1) our SLEMD algorithm, i.e., TAPM at level-0, with the simplistic detector that uses a single keyframe and multiple keyframes and 2) multilevel TAPM with the SLEMD method. We also analyze the effect of the multilevel temporal matching procedure introduced in Section 4.2 and propose a method to speed up TAPM without sacrificing classification accuracy. In addition, we compare our midlevel CS feature with three low-level features.

We chose the following 10 events from the LSCOM lexicon [15], [16]: *Car Crash*, *Demonstration Or Protest*, *Election Campaign Greeting*, *Exiting Car*, *Ground Combat*, *People Marching*, *Riot*, *Running*, *Shooting*, and *Walking*. They are chosen because 1) these events have relatively higher occurrence frequency in the TRECVID data set [16] and 2) intuitively, they may be recognized from visual cues. The number of positive samples for each event class ranges from 54 to 877. We also construct a background class (containing 3,371 video clips) that does not overlap with the above 10 events. When training the SVM, the negative samples for each event comprise the video clips from the other nine events and the background class. We randomly choose 60 percent of the data for training and use the remaining 40 percent for testing. Since it is computationally prohibitive to compute the similarities among video clips and train multiple SVMs with cross validation over multiple random training and testing splits, we reported the results from one split. For a performance metric, we use noninterpolated Average Precision (AP) [43], [44], which has been used as the official performance metric in TRECVID. It corresponds to the multipoint AP value of a precision-recall curve and incorporates the effect of recall when AP is computed over the entire classification result set. We also define Mean AP (MAP) as the mean of APs over 10 events.

5.1 Data Set Description and Annotation

We provide additional information about the data set and the selected event classes in this section. The TRECVID video corpus is probably the largest annotated video benchmark data set available to researchers today. Through the LSCOM effort, 449 semantic concepts have been manually annotated by a large student group to describe the visual content in each of the 61,901 subshots. The video data includes 137 video programs from six broadcast sources (English, Arabic, and Chinese), covering realistic and diverse visual content. More than 50 groups participated in TRECVID 2006 and the top results on high-level feature extraction were from Tsinghua University [30], the IBM T.J. Watson Research Center [29], CMU [31], Columbia University [34], and UCF [32]. The best MAP over 20 selected concepts is 19.2 percent [43].

TABLE 1
Average Precision (%) at Different Levels of TAPM

Event Name	KF-CS	MF-CS	L0 (SLEMD)	L1	L2	L0+L1	L0+L1+L2	L0+L1+L2-d
Car Crash	10.5	26.9	39.2	49.2	50.5	49.2	***51.1***	51.0
Demonstration Or Protest	16.0	18.0	21.9	22.7	21.3	23.6	***23.6***	23.6
Election Campaign Greeting	7.2	8.7	12.7	12.2	12.7	13.4	***13.9***	13.7
Exiting Car	15.6	15.9	46.2	49.3	40.2	***51.4***	50.7	50.1
Ground Combat	38.8	40.0	41.9	43.9	43.3	43.9	***44.2***	44.1
People Marching	19.5	19.5	21.4	24.5	24.9	25.7	***25.8***	25.8
Riot	11.1	15.3	19.9	20.8	17.5	22.8	22.7	***22.9***
Running	78.9	82.1	85.7	83.9	85.3	86.6	***86.7***	86.6
Shooting	9.6	7.2	9.9	10.7	8.6	***12.2***	10.4	9.9
Walking	37.1	41.3	50.1	52.1	52.4	51.1	52.4	***52.8***
Mean AP	24.4	27.5	34.9	36.9	35.7	38.0	***38.2***	38.1

Note that the results of L0 are from SLEMD and the results of KF-CS and MF-CS are from the keyframe and multiframe-based algorithms using the CS features. The last row, referred to as Mean AP, is the mean of APs over 10 events.

Among the annotated concepts, 56 belong to the event/activity category, 10 of which are chosen as our evaluation set in this paper. The first version of the LSCOM annotation labels (available at [15]) were obtained by having the human annotators look at a keyframe for each subshot in order to improve throughput in the annotation process. Such a process is adequate for static concepts (e.g., indoor and people) but deficient for judging events that involve strong temporal attributes such as *Exiting Car*. Hence, in this work, we have conducted further steps to refine the event labels by having an annotator view all of the frames in a shot.[2]

The detailed annotation procedure is described as follows: Since we did not have the resources to conduct labeling over the entire corpus (about 80 hours) for the 10 concepts, we prefiltered the shots by using the original keyframe-based LSCOM annotations [15]. Only shots receiving a positive label in the keyframe-based annotations for a given event concept were reexamined. The annotator reviewed all of the extracted frames to judge the presence/absence, as well as the start/end boundaries, of the event. Starting with the shot-level temporal boundaries provided by NIST, we sampled each positive shot at a rate of 10 frames per second to extract image frames from each shot. A student who was not involved in the algorithmic design annotated each shot as positive or negative for a given event concept by visually inspecting the sequence of extracted frames. During this process, the annotator also identified the temporal start/end boundaries of the event, which may not coincide with the initial shot boundaries provided by TRECVID.

After refinement, there are 3,435 positive clips in total over the 10 events. We believe this benchmark set from broadcast news and our annotation effort nicely complement the existing video event evaluation sets, which are usually drawn from the surveillance and meeting room domains.

2. TRECVID merges a subshot shorter than two seconds with its previous subshot to form a shot. To be consistent with the TRECVID evaluation, we refine the event labels at the shot level.

5.2 Single-Level EMD versus Keyframe-Based Algorithm

We compare our SLEMD algorithm to the keyframe-based algorithm that has frequently been used in TRECVID [43], [44]. In this paper, we extract 108-dimensional CS features from all of the images whose elements are the decision values of independent SVM classifiers. These SVM classifiers are independently trained using each of the three low-level features mentioned in Section 2 for detecting the 36 LSCOM-lite concepts.[3]

The classical keyframe-based algorithm (referred to as "KF-CS") for event recognition only considers information from one keyframe for each video clip. Namely, they applied SVM on the kernel matrix computed between the keyframes of every two clips. The kernel matrix is calculated by applying the same exponential function as (3) to convert the euclidean distance computed from the low-level features or our midlevel CSs. We also report the results from a late fusion approach with a multiframe-based representation, referred to as "MF-CS" here. In MF-CS, we uniformly choose another four frames for each video clip and then use the same technique as in KF-CS to train four SVMs. Finally, we fuse the SVM decisions from four frames and keyframe using (7) with the equal weights.

The experimental results are shown in Table 1, from which we have the following observations: 1) MF-CS outperforms KF-CS in most cases, which demonstrates the importance of utilizing multiple frames instead of just a single keyframe for event recognition, and 2) the SLEMD algorithm that considers alignment among multiple frames achieves much better accuracy than the keyframe and multiframe-based algorithms KF-CS and MF-CS. When comparing SLEMD with MF-CS, the AP for the *Exiting Car* event class significantly increases from 15.9 percent to 46.2 percent.

3. The LSCOM-lite lexicon, a preliminary version of LSCOM, includes the 39 dominant visual concepts present in broadcast news videos, covering objects (e.g., car and flag), scenes (e.g., outdoor and waterscape), locations (e.g., office and studio), people (e.g., person, crowd, and military), events (e.g., people walking or running and people marching), and programs (e.g., weather and entertainment). Three of them overlap with our target concepts of events and thus are not used for our midlevel representations.

As mentioned in Section 3, it is possible to use other weights, e.g., unit weight 1 for SLEMD. Here, we compare SLEMD (also referred to as L0) with the default normalized weights $1/m$ and $1/n$ for two video clips, with L0-UnitWeight, in which the weights are fixed at 1. Our experiments demonstrate that L0 with normalized weights significantly outperforms L0-UnitWeight (34.9 percent versus 20.9 percent). When we compute the distances between any two subclips at level-l and level-2, we use the default normalized weights as well.

5.3 Multilevel Matching versus Single-Level EMD

We conduct experiments to analyze the effect of varying the number of levels used in the temporal alignment process (TAPM described in Section 4). Table 1 shows the results at three individual levels—level-0, level-1, and level-2, labeled as L0, L1, and L2, respectively. Note that TAPM at level-0 is equivalent to the SLEMD algorithm. We also report results using combinations of the first two levels (labeled as L0+L1) with uniform weights $h_0 = h_1 = 1$, as suggested in [18], [25]. To fuse all three levels, we use nonuniform weights $h_0 = h_1 = 1$ and $h_2 = 2$ (referred to as L0+L1+L2-d), which have been proposed in [18], [25], as well as uniform weights $h_0 = h_1 = h_2 = 1$ (labeled as L0+L1+L2).

In Table 1, we observe that methods fusing information from different levels, L0+L1 and L0+L1+L2, generally outperform the individual levels, demonstrating the contribution of our multilevel pyramid match algorithm. When comparing L0+L1+L2 with the level-0 algorithm, i.e., SLEMD, the MAP over 10 event classes is increased from 34.9 percent to 38.2 percent. Some event classes enjoy large performance gains, e.g., the AP for the *Car Crash* event increases from 39.2 percent to 51.1 percent. We observe that videos from *Car Crash* seem to have better defined and distinguished temporal structures, which was successfully captured by TAPM. In our experiments, L0+L1+L2 with equal weights is usually comparable to or slightly better than L0+L1+L2-d.

Table 1 also shows the contribution from each individual level. It is intuitive that the transition from the multiple-frame-based method to the SLEMD-based method (L0) produces the largest gain (27.5 percent to 34.9 percent). Adding information at level-1 (i.e., L1) helps in detecting almost all events, with the *Car Crash* class having the largest gain. Additional increases in the temporal resolution (i.e., L1 to L2) result in only marginal performance differences, with degradation even seen in some events (like *Exiting Car*, *Riot*, and *Shooting*). A possible explanation is that videos only contain a limited number of stages because the videos in our experiments are relatively short (about 44 frames per video clip). Therefore, the performance using finer temporal resolutions is not good. Though our multiresolution framework is general, at least for the current corpus, it is sufficient to include just the first few levels instead of much higher temporal resolutions. This is consistent with the findings reported in prior work using SPM for object recognition [18].

The results in Table 1 also indicate that there is no single level that is universally optimal for all different events. Therefore, a fusion approach combining information from multiple levels in a principled way is the best solution in practice.

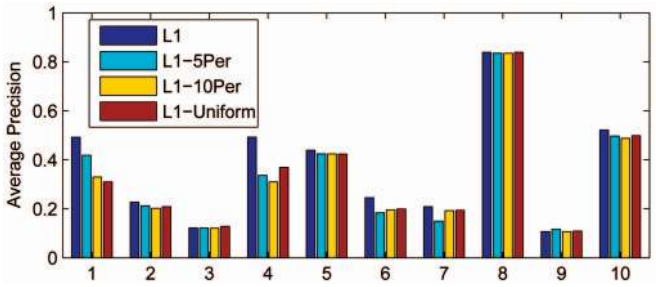

Fig. 5. Comparison with different temporal boundaries for TAPM at level-1. From left to right, the event classes are 1: car crash, 2: demonstration or protest, 3: election campaign greeting, 4: exiting car, 5: ground combat, 6: people marching, 7: riot, 8: running, 9: shooting, and 10: walking.

5.4 Sensitivity to Clustering Method and Boundary Precision

In Section 4.1, we applied T-HAC to obtain different subclips. We also take level-1 as an example to analyze the clustering results from T-HAC. For comparison, we consider two random perturbations of temporal boundaries, referred to as L1-5Per and L1-10Per here, and a uniform partition, denoted as L1-Uniform. Suppose the total number of frames in one video clip is n. For L1-Uniform, the temporal boundary is set as $Round(n/2)$, where $Round(x)$ is the nearest integer of x. The temporal boundaries of L1-5Per and L1-10Per are perturbed from T-HAC, i.e., we move the temporal boundary with the offset set as $Round(0.05n)$ and $Round(0.1n)$ frames, respectively, and the moving directions (forward or backward) randomly decided for different videos.

From Fig. 5, we have the following observations: 1) The results of L1 based on T-HAC are generally better than the results from other varied boundaries; therefore, T-HAC is used as a default method to generate the subclip boundaries in this work; and 2) for events *Car Crash*, *Exiting Car*, *People Marching*, and *Riot*, APs from L1 are significantly better. The possible explanation is that events like *Car Crash* and *Exiting Car* seem to have better defined and distinguished temporal structures. Temporal segmentation based on content clustering (as proposed T-HAC) and temporally aligned matching help in separating the distinct stages and matching their content.

5.5 The Effect of Temporal Alignment

To verify the effect of temporal alignment on detection performance, we also conducted experiments to evaluate an alternative method using temporal pyramid match without temporal alignment, referred to as L1-NoAlign and L2-NoAlign for level-1 and level-2, respectively. In such a detection method, subclips of one video are matched with subclips of the other video at the same temporal locations in each level. In other words, only values on the diagonal positions of the distance matrices shown in Figs. 3d and 3e are used in computing the distance between two video clips. The EMD process of finding the optimal flows among subclips is not applied. In Fig. 6, we observe that APs for the 10 events from L1-NoAlign and L2-NoAlign are generally worse than those of L1 and L2. Our experimental results

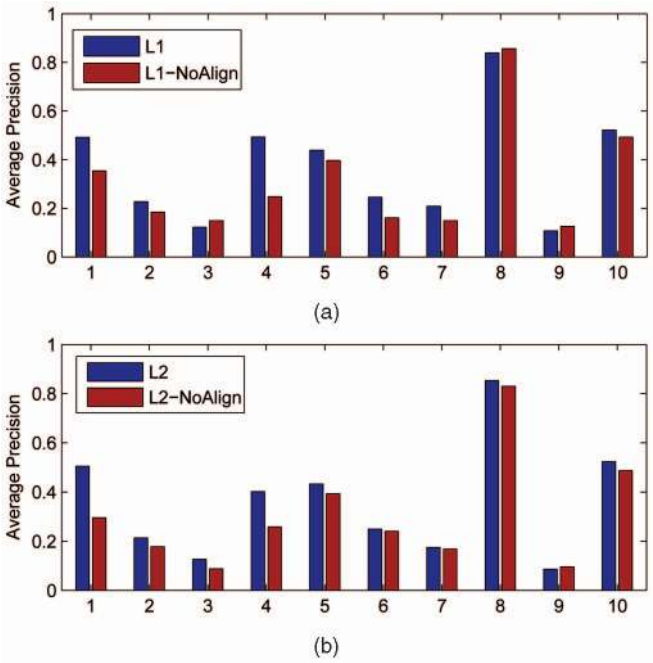

Fig. 6. Analysis of the temporal alignment algorithm. From left to right, the event classes are 1: car crash, 2: demonstration or protest, 3: election campaign greeting, 4: exiting car, 5: ground combat, 6: people marching, 7: riot, 8: running, 9: shooting, and 10: walking. (a) Results from TAPM at level-1. (b) Results from TAPM at level-2.

showed that such simplification significantly degrades the event detection performance, with 18.7 percent and 17.8 percent reductions in level-1 and level-2, respectively, in terms of MAP over 10 events. This confirms that temporal alignment in each level of the pyramid plays an important role. We also observe that L1 (or L2) significantly outperforms L1-NoAlign (or L2-NoAlign) for events *Car Crash* and *Exiting Car*, which have better defined and distinguished temporal stages than other events from the definition. The possible explanation is that broadcast newscasts are quite diverse. For example, videos of *Car Crash* may contain diverse scenes related to car crash and may not always follow certain temporal patterns.

Our results also show that a large portion of the optimal flow values (right-hand-side matrices in Figs. 3d and 3e) are indeed located at nondiagonal positions. About 50 percent of the optimal flows in level-1 and 90 percent in level-2 appear at nondiagonal locations, as shown in Table 2.

5.6 Algorithmic Complexity Analysis and Speedup

We analyze the algorithmic complexity and the average processing time of TAPM. Suppose the total number of frames in two video clips are the same, i.e., $m = n$, then the complexity of EMD is $O(m^3 log(m))$ [21]. If we uniformly partition video clips into subclips, the complexity involved in computing level-l distance is $O((2^{-l}m)^3 log(2^{-l}m))$. Therefore, TAPM at level-0 is the most time-consuming level.

TABLE 2
Percentage (%) of Optimal Flow Values Located at Nondiagonal Positions for Level-1 and Level-2

Algorithm	1	2	3	4	5	6	7	8	9	10
L1	50.5	49.0	43.2	45.6	48.6	47.7	49.6	48.4	49.3	48.0
L2	87.9	90.2	88.3	88.9	90.0	90.2	91.7	92.2	90.7	91.0

From left to right, the event classes are 1: car crash, 2: demonstration or protest, 3: election campaign greeting, 4: exiting car, 5: ground combat, 6: people marching, 7: riot, 8: running, 9: shooting, and 10: walking.

In our experiments, the total number of training samples is about 4,000. In Table 3, we report the average processing time for distance computation between one test sample with 4,000 training samples on a Pentium IV, 2.8 GHz PC with 1 Gbyte of RAM. In Table 3, we observe that the keyframe-based algorithm is very efficient because only one keyframe is used for decision, but its classification performance is poor, as demonstrated in Table 1. While the average processing times of L1 and L2 are significantly lower than L0, the speeds of L0, L1, and L2 are still much slower than the keyframe-based method. In the following, we describe one possible algorithm for speedup. We note that speed optimization is not the main focus of this paper and, thus, the proposed method is intended to demonstrate the feasibility of speed improvement, rather than achieving the fastest solution. Since TAPM at level-0 is the most time-consuming level, we take level-0 as an example. Note that our algorithm can also be used to speed up TAPM at level-1 and level-2.

Our solution is based on the observation that the algorithmic complexity of SLEMD depends on the total number of frames n in each video clip. One straightforward approach is simply to reduce the total number of frames by clustering. We test the idea by applying the K-Means algorithm to abstract the whole set of frames in video clips P and Q into several clusters. Then, the centroids of the ith cluster are used to define image sequences p_i and q_i and w_{p_i} and w_{q_i} are set as the relative size of the cluster (the number of frames in the cluster divided by the total number of frames in P and Q). Such reduced image sequences and weights can then be readily plugged into the same EMD procedure (i.e., (1) and (2)), with a computational complexity much less than that associated with the original image sequences. Note that the K-Means clustering method here is simply used to reduce the number of frames in each level without considering any temporal constraints. It is not intended to replace the temporally constrained clustering process T-HAC used to generate the subclip partitions needed in higher levels such as L1 and L2. As shown in Section 5.4, T-HAC plays an important role in obtaining

TABLE 3
Average Processing Time (Seconds) at Different Levels of TAPM

Algorithm	KF	L2	L1	L0	L0-KM20	L0-KM15	L0-KM10
Average Processing Time	0.002	13.693	16.225	20.972	1.700	0.940	0.416

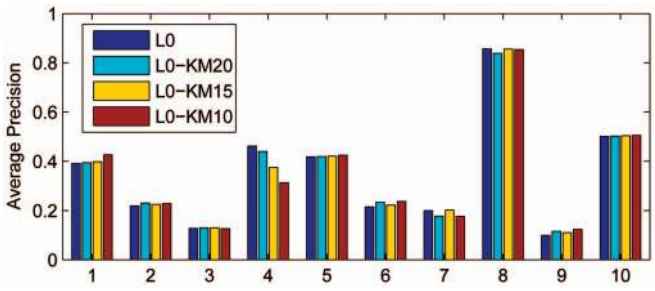

Fig. 7. Performance comparison of L0 with L0-KM20, L0-KM15, and L0-KM10. From left to right, the event classes are 1: car crash, 2: demonstration or protest, 3: election campaign greeting, 4: exiting car, 5: ground combat, 6: people marching, 7: riot, 8: running, 9: shooting, and 10: walking.

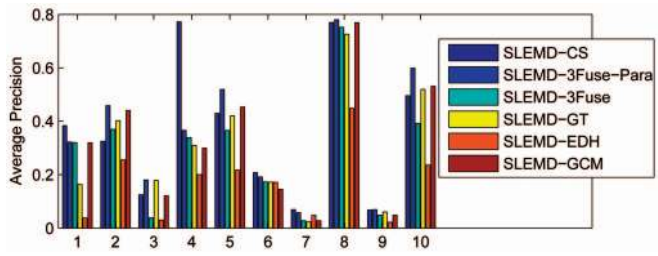

Fig. 8. Comparison of our midlevel concept feature (CS) with three low-level global features (GT-Gabor Texture, EDH-Edge Direction Histogram, and GCM-Grid Color Moment), as well as their fused results. From left to right, the event classes are 1: car crash, 2: demonstration or protest, 3: election campaign greeting, 4: exiting car, 5: ground combat, 6: people marching, 7: riot, 8: running, 9: shooting, and 10: walking.

adequate subclip boundaries and achieving the overall accuracies of event recognition.

In the K-Means clustering algorithm, we need to decide the total number of clusters beforehand. Considering that the videos in our experiments are relatively short, we investigate three cluster numbers: 20, 15, and 10. We denote the related results as L0-KM20, L0-KM15, and L0-KM10, respectively. In Fig. 7, we observe that the results from L0-KM10, L0-KM15, and L0-KM20 are comparable to the original TAPM at level-0 (except for a large performance drop for the *Exiting Car* event). Such results are interesting, indicating the temporal resolution required for event recognition (at least in the context of broadcast news) can be reduced to some extent without hurting the event detection accuracy. However, as shown in Table 3, when comparing temporally subsampled representations (L0-KM10, 15, 20) with L0, the average processing time is significantly reduced.

5.7 Concept Score Feature versus Low-Level Features

Finally, we compare the CS feature with three low-level global features for event recognition using SLEMD. Note that, in this comparison, the only difference is that low-level global features or our midlevel CS features are used to represent images. As introduced in Section 2, the low-level global features are Grid Color Moment, Gabor Texture, and Edge Direction Histogram [27], [28]. Considering that, we have used 90 video programs from the TRECVID 2005 data set for training the CS feature. For fair comparison, the test videos clips for event recognition are only from another 47 video programs in the TRECVID 2005 data set such that there is no overlap between the test data used in event recognition and the training data for training the CS feature. Note that the test video clips for event recognition used in this experiment are only part of the test data used in the previous experiments.

We report the fused results from three low-level features in a late fusion manner, referred to as "SLEMD-3Fuse" here by fusing the responses from three SVMs based on the above three low-level features using (7) with equal weights, which was also used in [33]. In contrast to early fusion techniques [20], [45], [46], which either concatenate input feature vectors or average multiple distances or kernels from different features to generate a single kernel, we choose late fusion techniques because of their successful use in TRECVID [29], [30], [31], [32], [34]. In addition, the same fusion technique as in (7) has been found to achieve some of the top performances in TRECVID by the IBM T.J. Watson Research Center [29].

We also run several different weights h_1, h_2, and h_3 to fuse three different features and then report the best results in the test data (referred to as "SLEMD-3Fuse-Para"). In this work, h_1, h_2, and h_3 varied from 0 to 1 with an interval 0.2, but we excluded a special case, $h_1 = h_2 = h_3 = 0$, which resulted in 215 combinations. While the parameter tuning technique mentioned above is not applicable in the real application, our motivation is to compare our midlevel CS feature with the best possible fused results from late fusion rather than developing new feature fusion algorithms. Feature fusion is still an open problem; please refer to [46] for recent advances.

We have the following observations from Fig. 8:

1. SLEMD-3Fuse is generally worse than SLEMD-GCM. A possible explanation is that the results from SLEMD-EDH are generally poor, which may degrade the performance of the equal weights-based fusion method SLEMD-3Fuse. We also observe that SLEMD-3Fuse-Para generally achieves better performance when compared with the single best feature, which demonstrates that it is possible to improve the performance by using feature fusion techniques.
2. Our CS feature generally outperforms the three low-level features, as well as SLEMD-3Fuse. MAPs over 10 events for SLEMD-CS, SLEMD-3Fuse, SLEMD-GT, SLEMD-EDH, and SLEMD-GCM are 36.5 percent, 28.3 percent, 29.8 percent, 16.7 percent, and 31.6 percent, respectively.
3. When comparing SLEMD-CS with SLEMD-3Fuse-Para, MAP also increases from 35.4 percent to 36.5 percent. In addition, SLEMD-3Fuse-Para uses the test data to choose the best parameter configuration, which is not practical in real applications.
4. Our CS feature is relatively robust for all 10 events, which demonstrates that CSs, fusing the decisions from 108 independent classifiers, effectively integrate information from multiple visual cues and abstract the low-level features into a robust midlevel representation.

6 Contributions and Conclusion

In this work, we study the problem of visual event recognition in unconstrained broadcast news videos. The diverse content and large variations in news video make it difficult to apply popular approaches using object tracking or spatiotemporal appearances. In contrast, we adopt simple holistic representations for each image frame and focus on novel temporal matching algorithms. We apply the SLEMD method to find optimal frame alignment in the temporal dimension and thereby compute the similarity between video clips. We show that the SLEMD-based temporal matching method outperforms the keyframe and multiframe-based classification methods by a large margin. Additionally, we propose TAPM to further improve event detection accuracy by fusing information from multiple temporal resolutions and explicitly utilizing the temporal information. We also discuss in depth our algorithm in terms of weight selection in SLEMD, sensitivity to temporal clustering, the effect of temporal alignment, and speedup. To the best of our knowledge, this work and our initial conference work are the first systematic studies of diverse visual event recognition in the unconstrained broadcast news domain with clear performance improvements.

We choose the scene-level CS feature in this work, which may not be sufficient to represent local characteristics. While the SIFT feature has been successfully used in object recognition, it is time consuming to extract SIFT features for a large data set. In the future, we will study other efficient and effective visual features, as well as other text and audio features. Currently, it is time consuming to compute the similarities among video clips and train multiple SVMs with cross validation over one training and testing split. While we have proposed using K-Means clustering method for speed-up, we will investigate other efficient algorithms to further reduce the computational complexity. Moreover, we will investigate effective techniques to fuse the information from different features, as well as multiple pyramid levels.

Acknowledgments

This material is based upon work funded by Singapore A*STAR SERC Grant (082 101 0018) and the US Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government. The authors also thank Akira Yanagawa and Wei Jiang for their help in generating the automatically detected concept scores from video and Tian Tsong Ng and Shuicheng Yan for their helpful discussions and suggestions. Part of this work was performed while Dong Xu was a postdoctoral research scientist at Columbia University.

References

[1] D. Zhang, D. Perez, S. Bengio, and I. McCowan, "Semi-Supervised Adapted HMMS for Unusual Event Detection," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 611-618, 2005.

[2] P. Peursum, S. Venkatesh, G. West, and H. Bui, "Object Labelling from Human Action Recognition," *Proc. IEEE Int'l Conf. Pervasive Computing and Comm.,* pp. 399-406, 2003.

[3] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 994-999, 1997.

[4] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 831-843, Aug. 2000.

[5] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury, "The Function Space of an Activity," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 959-968, 2006.

[6] C. Fanti, L. Manor, and P. Perona, "Hybrid Models for Human Motion Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1166-1173, 2005.

[7] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 462-469, 2005.

[8] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 166-173, 2005.

[9] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 726-733, 2003.

[10] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *Proc. IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 65-72, 2005.

[11] L. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 432-439, 2003.

[12] J. Niebles, H. Wang, and F. Li, "Unsupervised Learning of Human Action Categories Using Spatial Temporal Words," *Proc. British Machine Vision Conf.,* 2005.

[13] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. IEEE Int'l Conf. Pattern Recognition,* pp. 32-36, 2004.

[14] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Proc. Alvey Vision Conf.,* 1988.

[15] "Dto LSCOM Lexicon Definitions and Annotations," http://www.ee.columbia.edu/dvmm/lscom/, 2007.

[16] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-Scale Concept Ontology for Multimedia," *IEEE Multimedia,* vol. 13, no. 3, pp. 86-91, July-Sept. 2006.

[17] S. Ebadollahi, L. Xie, S.-F. Chang, and J.R. Smith, "Visual Event Detection Using Multi-Dimensional Concept Dynamics," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 881-884, 2006.

[18] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 2169-2178, 2006.

[19] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1470-1477, 2003.

[20] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *Int'l J. Computer Vision,* vol. 73, no. 2, pp. 213-238, June 2007.

[21] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l J. Computer Vision,* vol. 40, no. 2, pp. 99-121, Nov. 2000.

[22] R.M. Gray, D.L. Neuhoff, and P.C. Shields, "A Generalization of Ornstein's $\bar{d}$ Distance with Applications to Information Theory," *The Annals of Probability,* vol. 3, no. 2, pp. 315-328, Apr. 1975.

[23] E. Levina and P. Bickel, "The Earth Movers Distance Is the Mallows Distance: Some Insights from Statistics," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 251-256, 2001.

[24] S.T. Rachev, "The Monge-Kantorovich Mass Transference Problem and Its Stochastic Applications," *Theory of Probability and Its Applications,* vol. 29, pp. 647-676, 1984.

[25] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1458-1465, 2005.

[26] D. Xu and S.-F. Chang, "Visual Event Recognition in News Video Using Kernel Methods with Multi-Level Temporal Alignment," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[27] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M.R. Naphade, A. Natsev, J.R. Smith, J. Tesic, and T. Volkmer, "IBM Research TRECVID-2005 Video Retrieval System," *Proc. NIST TREC Video Retrieval Evaluation Workshop,* Nov. 2005.

[28] A. Yanagawa, W. Hsu, and S.-F. Chang, "Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors," technical report, Columbia Univ., July 2006.
[29] M. Campbell, A. Hauboldy, S. Ebadollahi, D. Joshi, M.R. Naphade, A. Natsev, J. Seidl, J.R. Smith, K. Scheinberg, J. Tesic, and L. Xie, "IBM Research TRECVID-2006 Video Retrieval System," *Proc. NIST TREC Video Retrieval Evaluation Workshop,* Nov. 2006.
[30] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z. Wang, Z. Xiang, J. Yuan, W. Zheng, B. Zhang, J. Zhang, L. Zhang, and X. Zhang, "Intelligent Multimedia Group of Tsinghua University at TRECVID 2006," *Proc. NIST TREC Video Retrieval Evaluation Workshop,* Nov. 2006.
[31] A.G. Hauptmann, M. Chen, M. Christel, D. Das, W.-H. Lin, R. Yan, J. Yang, G. Backfried, and X. Wu, "Multi-Lingual Broadcast News Retrieval," *Proc. NIST TREC Video Retrieval Evaluation Workshop,* Nov. 2006.
[32] J. Liu, Y. Zhai, A. Basharat, B. Orhan, S.M. Khan, H. Noor, P. Berkowitz, and M. Shah, "University of Central Florida at TRECVID 2006 High-Level Feature Extraction and Video Search," *Proc. NIST TREC Video Retrieval Evaluation Workshop,* Nov. 2006.
[33] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," technical report, Columbia Univ., Mar. 2007.
[34] "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," http://www.ee.columbia.edu/ln/dvmm/columbia374/, 2007.
[35] D. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1150-1157, 1999.
[36] P. Moreno, P. Ho, and N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications," *Proc. Neural Information Processing Systems,* Dec. 2003.
[37] F. Jing, M. Li, H. Zhang, and B. Zhang, "An Efficient and Effective Region-Based Image Retrieval Framework," *IEEE Trans. Image Processing,* vol. 13, no. 5, pp. 699-709, May 2004.
[38] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines,* http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2008.
[39] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, no. 3, pp. 264-323, Sept. 1999.
[40] P. Jensen and J. Bard, *Operations Research Models and Methods.* John Wiley & Sons, 2003.
[41] J. Munkres, "Algorithms for the Assignment and Transportation Problems," *J. Soc. for Industrial and Applied Math.,* vol. 5, no. 1, pp. 32-38, Mar. 1957.
[42] J.C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers,* 1999.
[43] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation Campaigns and TRECVID," *Proc. Eighth ACM Int'l Workshop Multimedia Information Retrieval,* pp. 321-330, 2006.
[44] "TRECVID," http://www-nlpir.nist.gov/projects/trecvid, 2008.
[45] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," *Proc. Int'l Conf. Image and Video Retrieval,* pp. 401-408, 2007.
[46] M. Varma and D. Ray, "Learning the Discriminative Power-Invariance Trade-Off," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

Dong Xu received the BEng and PhD degrees from the Electronic Engineering and Information Science Department at the University of Science and Technology of China in 2001 and 2005, respectively. During his PhD studies, he worked with Microsoft Research Asia and the Chinese University of Hong Kong. He also spent one year at Columbia University, New York, as a post-doctoral research scientist. He is currently an assistant professor at Nanyang Technological University, Singapore. His research interests include computer vision, pattern recognition, statistical learning, and multimedia content analysis. He is a member of the IEEE.

Shih-Fu Chang leads the Digital Video and Multimedia Lab in the Department of Electrical Engineering at Columbia University, conducting research in multimedia content analysis, image/video search, multimedia forgery detection, and biomolecular image informatics. Systems developed by his group have been widely used, including VisualSEEk, VideoQ, WebSEEk for visual search, TrustFoto for online image authentication, and WebClip for video editing. His group has received several best paper or student paper awards from the IEEE, ACM, and SPIE. His group has also made significant contributions to the development of MPEG-7 international multimedia standard. He is the editor-in-chief of *IEEE Signal Processing Magazine* (2006-2008), a recipient of a US Navy Office of Naval Research Young Investigator Award, an IBM Faculty Development Award, and a US National Science Foundation CAREER Award, and he has been a fellow of the IEEE since 2004. He has worked in different capacities at several media technology companies and served as a general cochair for the ACM Multimedia Conference 2000 and IEEE ICME 2004.