Visual Event Recognition in News Video using Kernel Methods with Multi-Level Temporal Alignment

Dong Xu and Shih-Fu Chang
Department of Electrical Engineering
Columbia University, New York, NY 10027
{dongxu, sfchang}@ee.columbia.edu

Abstract

In this work, we systematically study the problem of visual 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 Earth Mover's Distance (EMD) to integrate similarities among frames from two clips. Observing that a video clip is usually comprised of multiple sub-clips corresponding to event evolution over time, we further build a multi-level temporal pyramid. At each pyramid level, we integrate the information from different sub-clips with Integer-value-constrained EMD to explicitly align the sub-clips. 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 multi-level method clearly outperforms single-level EMD, and 2) single-level EMD outperforms by a large margin (43.0% in Mean Average Precision) basic detection methods that use only a single key-frame. Extensive analysis of the results also reveals an intuitive interpretation of subclip alignment at different levels.

1. Introduction

Event recognition from visual cues is a challenging task because of complex motion, cluttered backgrounds, occlusions, as well as geometric and photometric variances of objects. Previous work on video event recognition can be roughly classified as either activity recognition or abnormal event recognition. In model-based approaches to activity recognition, frequently-used models include HMM [17] and Dynamic Bayesian Network [16]. The work in [21] modeled each activity with a nominal activity trajectory and one function space for time warping. For model-based abnormal event recognition, Zhang et al. [22] used a semisupervised adaptive HMM framework. To model the relationship between different parts or regions, object tracking is usually performed before model learning [17]. Additionally these techniques heavily rely on the choice of good models, which in turn require sufficient training data to learn the model parameters.

Appearance-based techniques extract spatio-temporal features in the volumetric regions, which can be densely sampled or detected by salient region detection algorithms. For abnormal event recognition, Boiman and Irani [2] proposed to extract an ensemble of densely sampled local video patches to localize irregular behaviors in videos. For activity recognition, Ke et al. [10] applied boosting to choose volumetric features based on optical flow representations. [5] also used optical flow measurements in spatio-temporal volumetric regions. Other researchers extracted volumetric features from regions with significant local variations in both spatial and temporal dimensions [11, 15]. The performance of appearance-based techniques usually depends on reliable extraction of the spatial-temporal features and/or the 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 [10].

The extensive works mentioned above have demonstrated promise for event detection in domains such as surveillance or meeting room video. However, generalization to less constrained domains like broadcast news or consumer videos has not been demonstrated. Recently, due to the emerging applications of open-source intelligence and online video search, the research community has shown increasing interest in event recognition in broadcast news videos. Large-Scale Concept Ontology for Multimedia (LSCOM) [3] [14] 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

1-4244-1180-7/07/$25.00 ©2007 IEEE

including video analysts, knowledge representation experts, and video analytics researchers. Manual annotation of such event concepts have been completed for a large data set in TRECVID 2005 [3].

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 spatio-temporal interest regions, and robustly extract the spatial-temporal features.

To address the challenges of news video, Ebadollahi et al. [4] proposed to treat each frame in the video clip as an observation and apply HMM to model the temporal patterns of event evolution. Such approaches are distinct from most of the 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 [4] did not confirm clear performance improvement over a simplistic detection method using static information in key-frames 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 static representations for image frames without object tracking or spatio-temporal interest region detection. We propose a non-parametric approach in order to circumvent the problems of insufficient training data 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. To this end, we first represent one video clip as a bag of features, extracted from all the frames. To handle the temporal shift in different video clips, we apply the Earth Mover's Distance (EMD) [18], referred to as single-level EMD in this work, to measure video clip similarity, and combine it with SVM kernel classifiers for event detection. Single-level EMD computes the optimal flows between two sets of frames, yielding the optimal match between two frame sets, as shown in Fig. 1(a). With different settings of the feature weights (discussed in Sec. 2), EMD methods may be used to partially address the duration variation of events.

We also observe that one video clip is usually comprised of several sub-clips, which correspond to multiple stages of event evolution. For example, Fig. 2(a) and Fig. 2(f) 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 multi-stage structures, it makes sense to extend the single-level EMD mentioned above to multiple scales. Specifically, we propose a Temporally Aligned

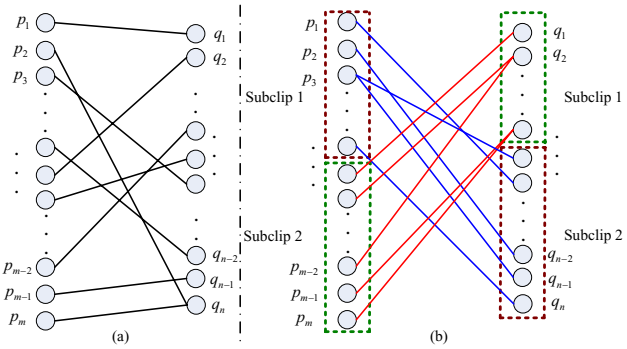

Figure 1. Illustration of matching in single-level EMD (a) and TAPM (b). For better viewing, please see the color pdf file.

Pyramid Matching (TAPM) framework, in which temporal matching is performed in a multi-resolution fashion. As shown in Fig. 1(b), frames in a sub-clip (conceptually corresponding to an event stage) are matched to frames in the same sub-clip in the other video, rather than spread over multiple sub-clips in Fig. 1(a). Such constraints may seem to be over-restrictive, but it explicitly utilizes the temporal information and is experimentally demonstrated to be effective for improving detection accuracy of several events in Section 4. 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 to fuse the matching results from multiple temporal scales, in a way similar to that used in Spatial Pyramid Matching (SPM) [12] and Pyramid Match Kernel (PMK) [6] for object recognition. However, it is important to note that in TAPM, unlike the fixed block-to-block matching method used in SPM, the sub-clips at different temporal locations may be matched as shown in Fig. 2(b), which will be explained in detail in Section 3.2.

We conduct comprehensive experiments on the large TRECVID 2005 database, and the experiments demonstrate that 1) TAPM clearly outperforms single-level EMD and 2) single-level EMD outperforms basic detection methods that use only one key-frame. To the best of our knowledge, this is the first work to systematically study the problem of visual event recognition in broadcast news video without any specific parametric model and object tracking.

2. Single-level Earth Mover's Distance in The Temporal Domain

EMD has shown promising performance in several different applications, such as content based image retrieval [9, 18], texture classification and general object recognition [23]. In this section, we develop a single-level EMD, to efficiently utilize the information from multiple frames for event recognition. We will extend this method to multiple levels in the next section. 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 i-th frame, w_{p_i} is the weight of the i-th frame. The weight w_{p_i} is used as the total supply of suppliers or the total capacity of consumer in the EMD method, with the default value of $1/m$ [1]. p_i can be any feature, such as Grid Color Moment [1] or Gabor Texture [1]. 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. 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 [2]), and $\widehat{f}_{ij}$ is the optimal flow that can be determined by solving the following linear programming problem in Eq. (2). $\widehat{f}_{ij}$ can be interpreted as the optimal match among frames from two video clips, as shown in Fig. 1(a).

$$\widehat{f}_{ij} = arg\min_{f_{ij}} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}$$

$$s.t. \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}); \quad f_{ij} \geq 0;$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, 1 \leq i \leq m; \sum_{i=1}^{m} f_{ij} \leq w_{q_j}, 1 \leq j \leq n. \quad (2)$$

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 non-negativity, symmetry and the triangle inequality holds in this case [18]. Suppose the total number of frames in two video clips are the same, i.e., $m = n$, then the complexity of EMD is $O(m^3 log(m))$ [18].

We apply EMD [18] in the temporal dimension to efficiently handle temporal shifts and different numbers of frames in different video clips. Fig. 2 (a) and Fig. 2 (f) show all the frames of the two video clips P (or P^0) and Q (or Q^0) from the "riot" class. Fig. 2(b) shows two representive key-frames 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. 2(c), in which the brighter pixels indicate that the values at that position are larger. In Fig. 2(c), we also use the red circles to indicate the positions of two key-frames. From this, we can see that the flow (calculated from the EMD process) between these two key-frames is very small, confirming that a key-frame based representation is not sufficient for capturing characteristics over multiple frames of the same event class. In Fig. 2(c), we also use four green circles 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 information among frames is not utilized in this single-level EMD method. In the next section, we will propose a multi-resolution framework to partially preserve the proximity relations.

For classification, we use a Support Vector Machine (SVM) because of its good performance [6, 9, 12, 23]. For a two-class (e.g., "riot" vs. "others") case, the decision function for a test sample P has the following form:

$$g(P) = \sum_{o} \alpha_o y_o K(P, Q_o) - b, \quad (3)$$

where $K(P, Q_o)$ is the value of a kernel function for the training sample Q_o and the test sample P, y_o is the class label of Q_o ($+1$ or -1), α_o is the learnt weight of the training sample Q_o and b is the threshold parameter. The training samples with weight $\alpha_o > 0$ are called support vectors. We use the Gaussian function to incorporate EMD distance into the SVM framework:

$$K(P,Q) = exp(-\frac{1}{A}D(P,Q)). \quad (4)$$

We set hyper-parameter 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 [9][23], EMD-kernel works well in content based image retrieval and object recognition.

3. Temporally Aligned Pyramid Matching

We observe that one video clip is usually comprised of several sub-clips, which correspond to event evolution over multiple stages. For example, from Fig. 2(a) and (f), 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 Spatial Pyramid Matching [12] and Pyramid Match Kernel [6] have demonstrated that better results may be obtained by fusing the information from multiple resolutions according to the pyramid structure in the spatial domain and feature domain respectively. Inspired by their work, we propose to apply Temporal-constrained Hierarchical Agglomerative Clustering (T-HAC) to build a multi-level pyramid in the temporal domain. According to the multi-level pyramid structure, each video clip is divided into several sub-clips, which

[1] Our experiments demonstrate that EMD with the normalized value works better than other possible weights, e.g. unit weight 1. The same setting was also used in [9][23]. We note that we may sacrifice event duration invariance to some extent with this setting.

[2] While it is also possible to use other distances in EMD, we choose Euclidian distance because of its simplicity and successful use in [9][23].

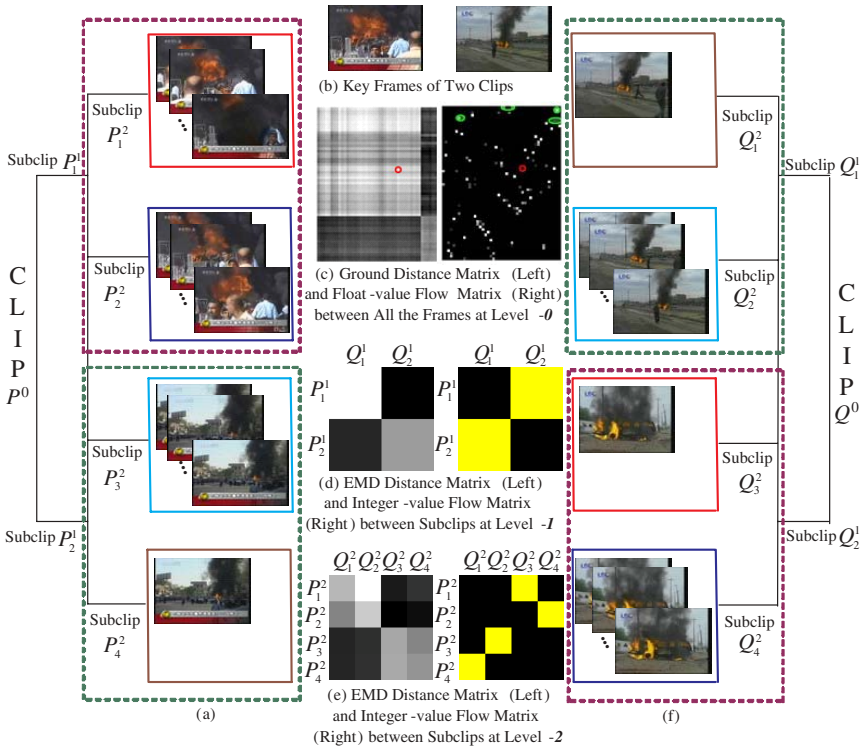

Figure 2. Conceptual illustration for Temporally Aligned Pyramid Matching. For better viewing, please see the color pdf file. (a)(f) Frames from two videos, P and Q, from the "riot" class are segmented into multiple sub-clips. (b) One key-frame for each clip. (c) Ground distance matrix and continuous-value flow matrix from the single-level EMD alignment process. The red circles indicate the locations of the key-frames 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)(e) The EMD distance matrix between the sub-clips and its 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 sub-clip of the same border color in Q, thus temporal proximity is preserved.

may represent an evolution stage 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 sub-clips at level-l as P_r^l, $r = 1, \ldots, R = 2^l$, $l = 0, \ldots, L-1$ with L as the total number of levels. For example, as shown in Fig. 2(a) and (f), four sub-clips at level-2 are denoted as P_1^2, P_2^2, P_3^2 and P_4^2, which are bounded with solid bounding box, and two sub-clips at level-1 are denoted as P_1^1 and P_2^1, which are bounded with a dashed bounding box.

We also observe that in broadcast news videos, event 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 sub-clips with Integer-value-constrained EMD to explicitly align the orderless sub-clips. Finally, we fuse the information from different levels of the pyramid, which results in the Temporally Aligned Pyramid Matching (TAPM).

3.1. Temporal Constrained Hierarchical Agglomerative Clustering

We use Hierarchical Agglomerative Clustering [7] to decompose a video into sub-clips. Clusters are constructed by iteratively combining existing clusters based on their distances. To incorporate temporal information, we propose to use Temporal Constrained Hierarchical Agglomerative Clustering (T-HAC) [3], in which at each step we only merge neighboring clusters in the temporal dimension. Examples of the clustering results are also shown in Fig. 2(a) and (f). It is obvious that T-HAC provides a reasonable pyramid structure in the temporal dimension. We also note that how to acquire the optimal clustering results is still an open problem, and other clustering algorithms may be used in our framework.

3.2. Alignment of Different Sub-clips

After we build the pyramid structure in the temporal dimension, for two video clips P and Q, we need to compute *level-l distance* $S^l(P, Q)$ between them. Firstly, we apply Eq. (1) to compute the EMD distance D^l. Fig. 2 (d) and (e) (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 cor-

[3] A simple solution for dividing video clips into different stages is to uniformly partition one clip into several sub-clips.

responding sub-clips. 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 uniformly partition video clips into the sub-clips, the complexity involved in computing level-l distance is $O((2^{-l}m)^3 log(2^{-l}m))$, which is significantly lower than the single-level EMD (i.e., $l = 0$). Therefore, fusing information from multiple levels does not add a significant computational cost to the overall detection method.

If we follow the same strategy as in Spatial Pyramid Matching [12], which assumes that the corresponding sub-regions 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, from Fig. 2(d), it is obvious that the diagonal elements of D^l are very large, because the first half stages of P and the last half stages of Q focus on the fire scene, and the last half stages of P and the first half stages of Q focus more on smoke and the nature scene. Thus it is desirable to align the subclips inversely in the temporal domain. Similar observations can be found from the EMD distance at level-2 in Fig. 2(e).

To explicitly align the different sub-clips 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 standard tools (e.g., simplex method from Matlab) for Linear Programming, according to the following Theorem 1:

Theorem 1 ([8]) *For the Linear Programming problem,*

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

$$\sum_{c} F_{rc} = 1, \ \forall r; \quad \sum_{r} F_{rc} = 1, \ \forall c; \quad \text{and} \quad R = C, \qquad (5)$$

it will always have an integer optimum solution when solved with the simplex method.

More details about the theorem can be found in [8]. We use the integer-value constraint to explicitly enforce the constraint that neighboring frames from a sub-clip are mapped to neighboring frames in the same sub-clip in the other video. Without such an integer-value constraint, one sub-clip may be matched to several sub-clips, resulting in a problem that one frame is mapped to multiple distant frames (as highlighted by the green circles in Fig. 2(c)). Note that non-integer mappings are still permitted at level-0 so that soft temporal alignments of frames are still allowed within a sub-clip. Similar to Eq. (1), *level-l distance* $S^l(P, Q)$ between two clips P and Q can be computed from integer-value flow matrix F^l and the distance matrix D^l by

$$S^l(P, Q) = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} F^l_{rc} D^l_{rc}}{\sum_{r=1}^{R} \sum_{c=1}^{C} F^l_{rc}} \qquad (6)$$

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

Again, we take the sub-clips at level-1 as an example. According to Theorem 1, we set $R = C = 2$, and obtain a 2×2 integer flow matrix F^1_{rc}, as shown in Fig. 2(d) (the right-hand-side matrix). From it, we can find that two sub-clips 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 of Fig. 2(e), where we set $R = C = 4$. In Fig. 2 (a) and (f), the aligned sub-clips are shown with the same border color.

For each level l, with the level-l distances $S^l(P, Q)$, we also apply Eq. (4) to compute 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 Eq. (3), 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.

3.3. Fusion of Information from Different Levels

As shown in [12], 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:

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

where h_l is the weight for level-l.

In our experiments, we set $L = 3$. We tried two weights: 1) equal weights, $h_0 = h_1 = h_2 = 1$, and 2) the weights suggested in [6, 12], i.e., $h_0 = h_1 = 1$, $h_2 = 2$. The experiments demonstrate that the results with equal weights are comparable to or slightly better than the weights suggested in [6, 12].

4. Experiments

We conduct experiments over the large TRECVID 2005 video corpus to compare 1) our single-level EMD algorithm, i.e., TAPM at level-0, with the simplistic detector that uses a key-frame only; 2) multi-level TAPM with the single-level EMD method. In addition, we also compare TAPM with temporal pyramid matching without alignment. We chose the following ten events from the LSCOM lexicon [3] [14]: *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 had relatively higher occurrence frequency in the TRECVID data set [14]; 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), which does not overlap the above

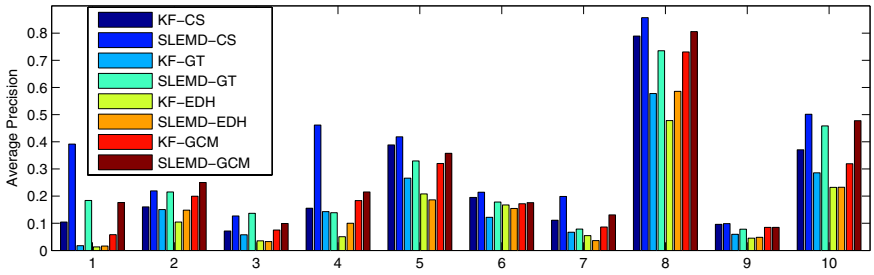

Figure 3. Comparison of our Single-level EMD (SLEMD) with the classical Key-Frame based algorithm on four different features (CS-concept score, GT-Gabor Texture, EDH-Edge Direction Histogram, and GCM-Grid Color Moment). From left to right, the event classes are 1: Car Crash, 2: Demonstration Or Protest, 3: Election Campaign Greeting, 4: Exiting Car, 5: Ground Combat, 6: People Marching, 7: Riot, 8: Running, 9: Shooting and 10: Walking.

Table 1. Average Precision (%) at different levels of TAPM. Note that the results of L0 are from Single-level EMD (SLEMD) and the results of KF-CS are from the key-frame based algorithm using the concept score features. The last row, referred as Mean AP, is the mean of APs over ten events.

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

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% of the data for training and use the remaining 40% for testing. For a performance metric, we use non-interpolated Average Precision (AP) [20], which has been used as the official performance metric in TRECVID. It corresponds to the area under the (non-interpolated) recall/precision curve, and incorporates the effect of recall when AP is computed over the entire classification result set.

4.1. 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. 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 [3]) were obtained by having the human annotators look at a key-frame for each subshot, in order to improve throughput in the annotation process. Such a process is adequate for static concepts (e.g., indoor, people), but deficient for judging event labels. Hence, in this paper, we have conducted further steps to refine the event labels by having an annotator view all the frames in a shot [4] in order to judge the presence of the event and obtain the precise start/end boundaries of the event. We will make the refined event annotations (3,435 positive clips in total over 10 events) publicly available. 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.

4.2. Single-level EMD vs. Key-frame based

First, we compare our single-level EMD algorithm to the key-frame based algorithm that has been frequently used in

[4] TRECVID merges a subshot shorter than 2 seconds with its previous subshot to form a shot. We refine the event labels at the shot level.

TRECVID [20]. We consider three low-level global features. For the Grid Color Moment (GCM) feature [1], we extract the first 3 moments of 3 channels of the CIE Luv color space over 5×5 fixed grid partitions, and aggregate the features into a single 225-dimensional feature vector. For the Gabor Texture (GT) feature [1], we take 4 scales and 6 orientations of Gabor transformations and further use their means and standard deviations, giving a dimension of 48. For the Edge Direction Histogram (EDH) feature [1], we extract 73 dimensions with 72 bins for edge direction quantized at 5 degrees and one bin for non-edge points.

In addition to the low-level features, we also include a mid-level representation, in which each image is mapped to a vector in a high-dimensional semantic space. Each element of the semantic vector represents the confidence score produced by a semantic concept classifier. Such a mid-level representation has shown promise for abstracting visual content [1]. In this paper, we use a 108-dimensional Concept Score (CS) vector, 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 above for detecting the 36 LSCOM-lite concepts. [5]

While it is possible to use other local features, such as tf-idf features [19] based on SIFT local descriptors [13], we use the above global features and scene-level concept scores because: 1) they can be efficiently extracted over the large video corpus; 2) they have been shown effective for detecting several concepts in previous TRECVID experiments [1]; 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. 2, and the experimental results shown in this section also confirm the potential of the representations.

The classical key-frame based algorithm [1] for event recognition only considers information from the key-frames. Namely, they applied SVM on the kernel matrix computed between the key-frames of every two clips. The experimental results, shown in Fig. 3, confirm that the single-level EMD algorithm that considers alignment among multiple frames achieves much better accuracy than the key-frame based algorithm. Using the concept score feature, the relative improvement by the single-level EMD method can be as high as 273.3% for some concepts, e.g., AP for the "car crash" event class increases from 10.5% to 39.2%. The relative improvements of Mean Average Precision (MAP), which is the mean of APs over ten events, are 43.0%, 44.6%, 10.8% and 24.2%, using the concept score, GT, EDH, or GCM features respectively. These numbers also confirm that the performance from the concept score feature is significantly better than that using the other three low-level features. A possible explanation is that concept scores, fusing the decisions from 108 independent classifiers, effectively integrate information from multiple visual cues and abstract the low-level features into a robust mid-level representation. In the following experiments, we only use the concept score feature.

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

4.3. Multi-level Matching vs. Single-Level EMD

We conduct experiments to compare TAPM at different levels using the concept score feature. Table I shows the results at three individual levels - level-0, level-1 and level-2, labeled as L0, L1, L2 respectively. Note that TAMP at level-0 is equivalent to the single-level EMD algorithm. We also report results using combinations of the first two levels (labeled as L0+L1) and all three levels (labeled as L0+L1+L2) with uniform weights. In addition, we experimented with non-uniform weights, $h_0 = h_1 = 1$, and $h_2 = 2$, which have been proposed in [6, 12] for fusing multiple scales in pyramid matching. We label such a combination method as L0+L1+L2-d.

From Table I, it is obvious that methods fusing information from different levels, L0+L1 and L0+L1+L2 generally outperform the individual levels, demonstrating the contribution of our multi-level pyramid match algorithm. When comparing L0+L1+L2 with the level-0 algorithm, i.e., single-level EMD, the MAP over ten event classes is increased from 34.9% to 38.2%, equivalent to a 9.5% relative improvement. Some event classes enjoy large performance gains, e.g., the AP for the "car crash" event increases from 39.2% to 51.1%. In our experiments, L0+L1+L2 with equal weights is usually comparable to or slightly better than L0+L1+L2-d.

Table I also shows the contribution from each individual level. It is intuitive that the transition from the key-frame based method to the single-level EMD based method (L0) produces the largest gain (24.4% to 34.9%). 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"). The possible explanation is that videos only contain limited stages because the videos in our experiments are relatively short. So the performance using finer temporal resolutions is not good. Though our multi-resolution 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 con-

sistent with the findings reported in prior work using spatial pyramid matching for object recognition [12].

The results in Table I 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.

4.4. 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 at level-1 and level-2. In such a detection method, sub-clips of one video are matched with sub-clips 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 Fig. 2 (d) and (e) are used in computing the distance between two video clips. The process of finding the optimal flows among sub-clips is not applied. Our experimental results showed that such simplification significantly degrades the event detection performance, with 18.7% and 17.8% reductions in level-1 and level-2 respectively in terms of MAP over ten events. This confirms that temporal alignment in each level of the pyramid plays an important role.

5. 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 spatio-temporal appearances. In contrast, we adopt simple holistic representations for each image frame and focus on novel temporal matching algorithms. We apply the single-level EMD method to find optimal frame alignment in the temporal dimension and thereby compute the similarity between video clips. We show that the mid-level representation based on semantic concepts produces a significantly superior accuracy compared to classical image-level features. We also show that the single-level EMD based temporal matching method outperforms the key-frame based classification method by a large margin. Additionally, we propose Temporally Aligned Pyramid Matching to further improve event detection accuracy by fusing information from multiple temporal resolutions and explicitly utilizing the temporal information. To the best of our knowledge, this work represents the first systematic study of diverse visual event recognition in the unconstrained broadcast news domain, with clear performance improvements.

6. Acknowledgment

This work was funded in whole by the U.S. Government VACE program. The views and conclusions are those of the authors, not of the U.S. Government or its agencies.

References

[1] A. Amir et al , *IBM Research TRECVID-2005 Video Retrieval System*, In NIST TRECVID Workshop, 2005. 3, 7

[2] O. Boiman and M. Irani, *Detecting Irregularity in Images and in Video*, Proceedings of ICCV, 2005. 1

[3] DTO LSCOM Lexicon Definitions and Annotations, *http://www.ee.columbia.edu/dvmm/lscom/*. 1, 2, 5, 6

[4] S. Ebadollahi et al, *Visual Event Detection Using Multi-Dimensional Concept Dynamics*, Proceedings of ICME, 2006. 2

[5] A. Efros, A. Berg, G. Mori and J. Malik, *Recognizing Action at a Distance*, Proceedings of ICCV, 2003. 1

[6] K. Grauman and T. Darrell, *The Pyramid Match Kernel: Discriminantive Classification with Sets of Image Features*, Proceedings of ICCV, 2005. 2, 3, 5, 7

[7] A. Jain, M. Murty and P. Flynn, *Data Clustering: A Review*, ACM Computing Surveys, 1999. 4

[8] P. Jensen and J. Bard, *Operations Research Models and Methods*, John Wiley and Sons, 2003. 5

[9] F. Jing et al, *An Efficient and Effective Region-based Image Retrieval Framework*, IEEE Transactions on Image Processing, 2004. 2, 3

[10] Y. Ke, R. Sukthankar and M. Hebert, *Efficient Visual Event Detection Using Volumetric Features*, Proceedings of ICCV, 2005. 1

[11] L. Laptev and T. Lindeberg, *Space-time Interest Points*, Proceedings of ICCV, 2003. 1

[12] S. Lazebnik, C. Schmid and J. Ponce, *Beyond Bags of Features, Spatial Pyramid Matching for Recognizing Natural Scene Categories*, Proceedings of CVPR, 2006. 2, 3, 5, 7, 8

[13] D. Lowe, *Object Recognition from Local Scale-Invariant Features*, Proceedings of ICCV, 1999. 7

[14] M. Naphade et al, *Large-Scale Concept Ontology for Multimedia*, IEEE Multimedia Magazine, 2006. 1, 5

[15] J. Niebles, H. Wang and F. Li, *Unsupervised Learning of Human Action Categories Using Spatial Temporal Words*, Proceedings of BMVC, 2005. 1

[16] N. Oliver, B. Rosario and A. Pentland, *A Bayesian Computer Vision System for Modeling Human Interactions*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. 1

[17] P. Peursum et al, *Object Labelling from Human Action Recognition*, Proceedings of PerCom, 2003. 1

[18] Y. Rubner, C. Tomasi and L. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, International Jouranl of Computer Vision, 2000. 2, 3

[19] J. Sivic and A. Zisserman, *Video Google: A Text Retrieval Approach to Object Matching in Videos*, Proceedings of ICCV, 2003. 7

[20] TRECVID, *http://www-nlpir.nist.gov/projects/trecvid*. 6, 7

[21] A. Veeraraghavan, R. Chellappa and A.K. Roy-Chowdhury, *The Function Space of an Activity*, Proceedings of CVPR, 2006. 1

[22] D. Zhang et al, *Semi-supervised Adapted HMMs for Unusual Event Detection*, Proceedings of CVPR, 2005. 1

[23] J. Zhang et al, *Local Features and Kernels for Classification of Texture and Object Categories: An In-Depth Study*, International Journal of Computer Vision, 2007. 2, 3