Event Recognition in Videos by Learning from Consumer Videos

#**K.Iniya**[1] M.Tech (Multimedia Technology), Computer Science and Engineering
#**G.S.Rizwana Banu**[2], Assistant Professor, Computer Science and Engineering
# KSR College of Engineering, Namakkal, Tamil Nadu, INDIA.

Abstract: A large number of loosely labelled web videos (e.g., from YouTube) and web images (e.g., from Google/Bing image search) are leveraged for visual event recognition in consumer videos without requiring any labeled consumer videos. This task is formulated as a new multi-domain adaptation problem with heterogeneous sources, in which the samples from different source domains can be represented by different types of features with different dimensions (e.g., the SIFT features from web images and space-time (ST) features from web videos) while the target domain samples have all types of features, we propose a new transfer learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL), in order to 1) fuse the information from multiple pyramid levels and features (i.e., space-time features and static SIFT features) and 2) cope with the considerable variation in feature distributions between videos from two domains (i.e., web video domain and consumer video domain). For each pyramid level and each type of local features, we first train a set of SVM classifiers based on the combined training set from two domains by using multiple base kernels from different kernel types and parameters, which are then fused with equal weights to obtain a pre learned average classifier. In A-MKL, for each event class we learn an adapted target classifier based on multiple base kernels and the pre learned average classifiers from this event class or all the event classes by minimizing both the structural risk functional and the mismatch between data distributions of two domains. Extensive experiments demonstrate the effectiveness of our proposed framework that requires only a small number of labelled consumer videos by leveraging web data.

Keywords: Event Recognition, transfer learning, domain adaptation, adaptive MKL, aligned space-time pyramid matching.

1.INTRODUCTION

This Action recognition “in the wild” is often a very difficult problem for computer vision. When the camera is non-stationary, and the background is fairly complicated, it is often difficult to infer the foreground features and the complex dynamics that are related to an action. Moreover, motion blur, serious occlusions and low resolution present additional challenges that cause the extracted features to be largely noisy. Under such challenges, we argue that the features of the scene and/or moving objects can be used to complement features extracted from people in the video. In this work, our aim is to capture such relationships between objects, scenes and actions. Our approach starts with extracting a large set of features for describing both the shape and the motion information in the videos. All the features are extracted densely, allowing spatial and temporal overlap, and we operate over tracks when the temporal continuity is available. We do not use any explicit object detectors, but treat each moving region as an object candidate. In the end, the videos are represented with multiple feature vectors acquired from different feature channels. We are particularly interested in human action classification in the real-world, i.e. in unconstrained video sources like YouTube. Moreover, there may be more than one person or moving object in the video, and only a subset of the detected regions are involved in the action. Our aim is to be able to train our action models in the presence of such diverse conditions. Our proposed framework is shown in Fig. 2 and consists of two contributions. First, we extend the recent work on pyramid matching [1], [4], [5], [12], [13] and present a new matching method, called Aligned Space-Time Pyramid Matching (ASTPM), to effectively measure the distances between two video clips that may be from different domains. Specifically, we divide each video clip into space-time volumes over multiple levels. We calculate the pair-wise distances between any two volumes and further integrate the information from different volumes with Integer-flow Earth Mover’s Distance (EMD) to explicitly align the volumes. The second is our main contribution. For this purpose, formulate a new transfer learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL).

Specifically, we first obtain one pre-learned classifier for each event class at each pyramid level and with each type of local feature, in which existing kernel methods can be readily employed. In this work, we adopt the pre-learned average classifier by equally fusing a set of SVM classifiers that are pre-learned based on a combined training set from two domains by using multiple base kernels from different kernel types and parameters. For each event class, we then learn an adapted classifier based on multiple base kernels and the pre-learned average classifiers from this event class or all event classes by minimizing both the structural risk functional and mismatch between data distributions of two domains. We test our approach over the extensive YouTube dataset provided by Liu et al [3], and the results demonstrate that the proposed framework effectively combines different and noisy feature channels for accurate human action recognition.

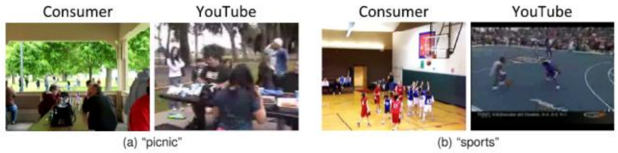


Fig.1. Four sample frames from consumer videos and YouTube videos. Our work aims to recognize the events in consumer videos by using a limited number of labeled consumer videos and a large number of YouTube videos. The examples from two events (i.e.,"picnic" and "sports") illustrate the considerable appearance differences between consumer videos and YouTube videos, which poses great challenges to conventional learning schemes but can be effectively handled by our transfer learning method A-MKL.

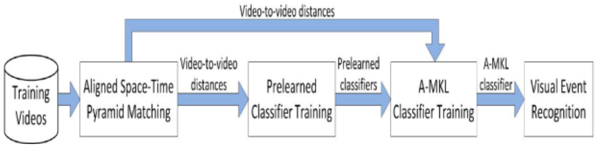


Fig.2. The flowchart visual event recognition framework. It consists of an aligned space-time pyramid matching method that effectively measures the distances between two video clips and a transfer learning method that effectively copes with the considerable variation in feature distributions between the web videos and consumer videos.

2.RELATED WORK

Human action recognition has been a very active research topic over the recent years. This makes the comprehensive listing of the related literature impossible, while presents an extensive review of the subject. Some of the recent works include In most of the earlier works, the focus is on simpler scenarios, where the background was stable and the foreground human figure is easy to extract. However, this scenario is hardly realistic; videos from the real world are fairly complicated, especially when taken in uncontrolled environments. Some recent approaches try to deal with such complex scenarios .Joint modelling of object and action interactions has been a recent topic of interest. Moore et al.'s work is one of the earliest attempts to consider actions and objects together. They use belief networks for modelling object and hand movements extracted from static camera sequences. try to improve the localization of both objects and actions by using a graphical Bayesian model, .use movie scripts as automatic supervision for scene and action recognition in movies. They assume that the objects related to each action are known beforehand and corresponding object detectors are available. In our case, we do not rely on explicit object detectors and try to discover related objects in an unsupervised manner. We consider each moving region as a candidate object region and we utilize shape and motion descriptors for all candidate regions. While doing this, we have no explicit knowledge about their class membership.

3.ALIGNED SPACE TIME PYRAMID MATCHING

Pyramid matching algorithms were proposed for different applications, such as object recognition, scene classification, and event recognition in movies and news videos [1], [4], [5], [12], [13]. These methods involved pyramidal binning in different domains (e.g., feature, spatial, or temporal domain), and improved performances were reported by fusing the information from multiple pyramid levels. Spatial pyramid matching [5] and its space-time extension [4] used fixed block-to-block matching and fixed volume-to-volume matching (we refer to it as unaligned space time matching), respectively. In contrast, our proposed Aligned Space-Time Pyramid Matching extends the methods of Spatially Aligned Pyramid Matching (SAPM) [12] and Temporally Aligned Pyramid Matching (TAPM) [13] from either the spatial domain or the temporal domain to the joint space-time domain, where the volumes across different space and time locations can be matched.

Similarly to [4], we divide each video clip into 8^l non-overlapped space-time volumes over multiple levels, l= 0, . . . , L - 1, where the volume size is set as $1/2^l$ of the original video in width, height, and temporal dimension. Fig.3 illustrates the partitions of two videos Vi and Vj at level-1. Following [4], we extract the local space-time (ST) features, including Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF), which are further concatenated together to form lengthy feature vectors. We also sample each video clip to extract image frames and then extract static local SIFT features [8] from them.

Our method consists of two matching stages. In the first matching stage, we calculate the pair-wise distance D_{rc} between each two space-time volumes V_i (r) and V_j (c), where r, c = 1, . . . ,R, with R being the total number of volumes in a video. The space-time features are vector quantized into visual words and then each space-time volume is represented as a token-frequency feature. As suggested in [4], we use x^2 distance to measure the distance D_{rc}. Noting that each space-time volume consists of a set of image blocks, we also extract token-frequency features from each image block by vector quantizing the corresponding SIFT features into visual words. And based on the token-frequency features, as suggested in [13], the pair-wise distance D_{rc} between two volumes V_i (r) and V_j (c) is calculated by using EMD [9] as follows:

$$D_{rc} = \frac{\sum_{u=1}^{H} \sum_{v=1}^{I} \widehat{f}_{uv} d_{uv}}{\sum_{u=1}^{H} \sum_{v=1}^{I} \widehat{f}_{uv}},$$

where H, I are the numbers of image blocks in V_i (r), V_j (c), respectively, d_{uv} is the distance between two image blocks (Euclidean distance is used in this work), and f_{uv} is the optimal flow that can be obtained by solving the linear programming problem. In the second stage, we further integrate the information from different volumes by using integer-flow EMD to explicitly align the volumes.

Theorem 1 ([18]): The linear programming problem,

$$\widehat{F}_{rc} = \arg\min_{F_{rc}} \sum_{r=1}^{R} \sum_{c=1}^{R} F_{rc} D_{rc},$$

$$\text{s.t.} \sum_{c=1}^{R} F_{rc} = 1,\ \forall r, \sum_{r=1}^{R} F_{rc} = 1,\ \forall c,$$

will always have an integer optimal solution when solved by using the Simplex method. In the next section, we will propose a new transfer learning method to fuse the information from multiple pyramid levels and different types of features.

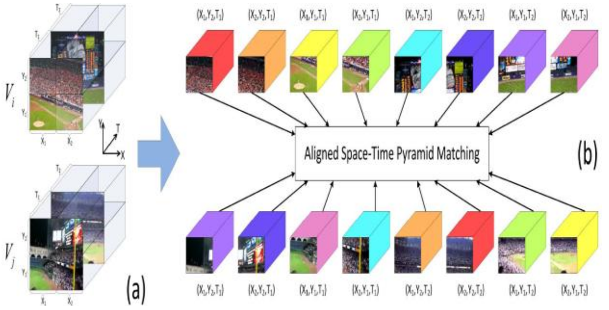


Fig.3. Illustration of the proposed Aligned Space-Time Pyramid Matching method at level-1: (a) each video is divided into eight space-time volumes along the width, height, and temporal dimensions. (b) The matching results are obtained by using our ASTPM method. Each pair of matched volumes from two videos is highlighted in the same colour.

4.ADAPTIVE MULTIPLE KERNEL LEARNING (AMKL)

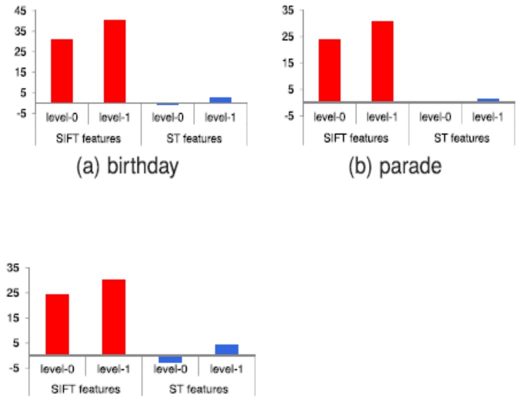
The New transfer learning method to learn a target classifier adapted from a set of pre-leared classifiers as well as a perturbation function that is based on multiple base kernels. In Adaptive SVM (ASVM) [14], the target classifier $f^T(x)$ is adapted from an existing classifier $f^A(x)$ (referred to as auxiliary classifier) trained based on the samples from the auxiliary domain. Specifically, the target decision function is defined as follows:

$$f^T(\mathbf{x}) = f^A(\mathbf{x}) + \Delta f(\mathbf{x}),$$

where Δf(x) is called a perturbation function that is learned by using the labeled data from the target domain only. The pre learned classifiers are used as prior for learning a robust adapted target classifier. In A-MKL, the existing machine learning methods (e.g., SVM, FR, and so on) using different types of features (e.g., SIFT and ST features) can be readily used to obtain the pre learned classifiers. Moreover, in contrast to A-SVM which uses the predefined weights to combine the pre learned auxiliary classifiers, we learn the linear combination coefficients of the pre learned classifiers in this work, where P is the total number of the pre learned classifiers. Specifically, we use the average classifiers from one event class or all the event classes as the pre learned classifiers. We additionally employ multiple predefined kernels to model the perturbation function in this work, because the utilization of multiple base kernels instead of a single kernel can further enhance the interpretability of the decision function and improve performances. We refer to our transfer learning method based on multiple base kernels as A-MKL because A-MKL can handle the distribution mismatch between the web video domain and the consumer video domain.

6.Experimental Results

In this work, events like, "birthday," "picnic," "parade," are chosen for experiments. In addition new consumer video clips from real users on our own are collected. Similarly to [7], new YouTube videos from the website are collected. Moreover, the consumer videos are annotated to determine whether a specific event occurred to watch each video clip rather than just look at the key frames.



(a) birthday

(b) parade

(c) picnic

7. Conclusion

Here we propose a new event recognition framework for consumer videos by leveraging a large amount of loosely labeled YouTube videos. Specifically, we propose a new pyramid matching method called ASTPM and a new transfer learning method, referred to as A-MKL, to better fuse the information from multiple pyramid levels and different types of local features and to cope with the mismatch between the feature distributions of consumer videos and web videos. This work cascade into the recent research trend of "Internet Vision," where the massive web data including images and videos together with rich and valuable contextual information (e.g., tags, categories, and captions) are employed for various computer vision and computer graphics applications such as image annotation [10], [11], image retrieval [6], scene completion [2], and so on.

8.References

[1] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features," Proc. 10th IEEE Int'l Conf. Computer Vision, pp. 1458-1465, 2005.

[2] J. Hays and A.A. Efros, "Scene Completion Using Millions of Photographs," ACM Trans. Graphics, vol. 26, no. 3, article 4, 2007.

[3] J.T. Kwok and I.W. Tsang, "Learning with Idealized Kernels," Proc. Int'l Conf. Machine Learning, pp. 400-407, 2003.

[4] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2008.

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

[6] Y. Liu, D. Xu, I.W. Tsang, and J. Luo, "Textual Query of Personal Photos Facilitated by Large-Scale Web Data," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 1022-1036, May 2011.

[7] A.C. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's Consumer Video Benchmark Data Set: Concept Definition and Annotation," Proc. Int'l Workshop Multimedia Information Retrieval, pp. 245-254, 2007.

[8] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int'l J. Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

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

[10] A. Torralba, R. Fergus, and W.T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1958-1970, Nov. 2008.

[11] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating Images by Mining Image Search Results," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1919-1932, Nov. 2008.

[12] D. Xu, T.J. Cham, S. Yan, L. Duan, and S.-F. Chang, "Near Duplicate Identification with Spatially Aligned Pyramid Matching," IEEE Trans. Circuits and Systems for Video Technology, vol. 20, no. 8, pp. 1068-1079, Aug. 2010.

[13] D. Xu and S.-F. Chang, "Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1985-1997, Nov. 2008.

[14] J. Yang, R. Yan, and A.G. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive SVMs," Proc. ACM Int'l Conf. Multimedia, pp. 188-197, 2007.