Learning Multiple Sequence-based Kernels for Video Concept Detection

Werner Bailer
JOANNEUM RESEARCH, DIGITAL – Institute for Information and Communication Technologies
Graz, Austria
Email: werner.bailer@joanneum.at

***Abstract*—Kernel based methods are widely applied to concept and event detection in video. Recently, kernels working on sequences of feature vectors of a video segment have been proposed for this problem, rather than treating feature vectors of individual frames independently. It has been shown that these sequence-based kernels (based e.g., on the dynamic time warping or edit distance paradigms) outperform methods working on single frames for concepts with inherently dynamic features. Existing work on sequence-based kernels either uses a single type of feature or a fixed combination of the feature vectors of each frame. However, different features (e.g., visual and audio features) may be sampled at different (possibly even irregular) rates, and the optimal alignment between the sequences of features may be different. Multiple kernel learning (MKL) has been applied to similarly structured problems, and we propose MKL for combining different sequence-based kernels on different features for video concept detection. We demonstrate the advantage of the proposed method with experiments on the TRECVID 2011 Semantic Indexing data set.**

***Keywords*-feature combination; fusion; learning**

I. Introduction

Recent research on concept detection in video increasingly considers dynamic concepts and events, making better use of the temporal dimension of video. This is evident both from the literature as well as from the increasing inclusion of dynamic concepts in the TRECVID [1] Semantic Indexing (SIN) task as well as the Multimedia Event Detection (MED) task introduced in 2010. This requires methods that are able to model concepts not only by samples from key frames, but by the sequence of samples of a video segment.

Sequence-based kernels, i.e., kernel functions that are able to determine the similarity of sequences of feature vectors, are one of the methods proposed for this purpose. They are commonly used in computationally biology, where they are often restricted to discrete features. Recently, appropriate sequence-based kernels supporting a wide range of features have been proposed for applications in multimedia, such as concept or event detection. Kernel-based machine learning methods, most notably Support Vector Machines (SVMs), have been successfully applied to concept detection in video. Experiments have shown that concept classifiers using sequence-based kernels outperform those using kernels matching only the individual feature vectors of the samples of a segment independently (see literature cited in Section I-B for detailed results).

The general approach of sequence-based kernels is to define a kernel function on a sequence[1] of feature vectors from two video segments (which may be regularly or irregularly sampled). Elements in the sequence represent the feature vectors of individual frames, and a base distance/similarity function (which can be a kernel itself) is applied to them. Then the kernel value for the two sequences is determined from the base distance similarity value, e.g., by choosing some optimal alignment, a weighted combination of different alignments etc. The latter step includes many properties that discriminate the different types of sequence-based kernels, such as thresholds for the base distance/similarity, constraints on gaps in the alignment, etc.

A. Motivation

The authors of [2] compared five different sequence-based kernels for concept detection. The kernels differ for example in whether they enforce a sequence in matching or allow elements to match across their order, and whether they consider a single optimal alignment or include several/all alignments in the result. The compared kernels yield quite similar results in terms of mean/median average precision over a set of 20 concepts from the TRECVID 2007 HLFE data set. However, for individual concepts there are considerable differences in the performance of the different types of kernels, and for some concepts, kernels treating the samples of a video segment independently perform best.

Existing work on sequence-based kernels either uses a single type of feature (e.g., bag of visual words) or combines the feature vectors of frames (e.g., by a weighted sum or product). When using multiple features, the optimal alignments between two sequences can vary in the different types of features. For example, for kernels supporting gaps in the alignment, a strong short-term lighting change might cause a gap in the alignment of a color feature, while a continuous alignment may still be possible for a texture based feature, thus increasing the value of the kernel function over the case where a gap is introduced for all features together. Also, audio and visual features may be extracted with different temporal sampling rates, so that they cannot be easily combined into feature vectors for a certain time point.

[1] In this paper, the term sequence denotes a possibly non-contiguous subsequence.

The optimal alignment determined by a sequence-based kernel also depends on parameters such as a similarity threshold for the values of the kernel function between individual elements in the sequence, the tolerable gap, or whether to base optimality on the length of the match or the mean similarity of the matching elements. Depending on the choice of these parameters, different alignments with associated values of the kernel function are possible, and it is often not possible to tell which of the alignments is "correct" or just "better" for a certain task. This consideration has led to the design of kernels that do not just consider a single alignment, but determine the kernel value as a (weighted) sum of several or in the extreme case *all* possible alignments. As it is difficult to determine the weights for the different alignments, they are often based on the same optimality criteria as in kernels choosing a single alignment, e.g. length of the matching sequence or they weighted equally.

Multiple kernel learning (MKL) has been proposed for problems, where instead of choosing a kernel a priori, weights for combining different kernels are learned together with the model [3]. In this paper, we propose MKL for combining different sequence-based kernels for video concept detection, with different parameterizations and using different features, as well as for combining sequence-based kernels with kernels treating the samples independently.

This paper is organized as follows. The rest of this section discusses relevant related work on sequence-based kernels and on multiple kernel learning. In Section II we propose multiple kernel learning for video concept detection using sequence-based kernels, and Section III presents experimental results on the TRECVID 2011 Semantic Indexing data set, a large scale set for concept detection. Section IV concludes the paper.

B. Related work

Several approaches based on the idea of the pyramid match kernel have been proposed. The original pyramid match kernel [4] partitions the feature space in each of the dimensions of the input feature vector. The approach has been extended to spatio-temporal matching in [5], using sets of clustered SIFT (scale-invariant feature transform) and optical flow features as local descriptors. The authors apply pyramid matching only to the image space, (i.e., subdividing an image into a spatial pyramid, and counting features of the same type in each of the bins) but use clustering in the feature space (i.e., the common bag of visual words approach). Another variant [6] uses temporally constrained hierarchical agglomerative clustering to build a structure of temporal segments. The similarity between segments is determined using the earth mover's distance (EMD) and the pyramid match kernel is applied to the similarities on the different hierarchy levels. The temporal order within the clips is aligned using linear programming.

The all subsequences kernel (ASS) [7] has been proposed for string matching, but can be applied to video sequences, determining the matching subsequences using dynamic programming. The authors of [8] propose a time sequence generalised alignment kernel (TSGA) that generalizes the idea of dynamic time warping by using the soft-max of all alignments rather than choosing a specific alignment. It is assumed that the kernel between two sequence elements is a conditionally positive kernel of the form $\kappa = e^{-d}$, with d being some distance.

The authors of [9] use the Levenshtein distance between sequences of clustered local descriptors for classification of still images. Recently, a kernel for event classification based on sequences of histograms of visual words has been proposed [10]. The authors consider different similarity measures between the histograms and use them instead of symbol equality in the Needleman-Wunsch distance and plug it into a Gaussian kernel. In [11] a kernel based on longest common subsequence (LCS) matching of sequences has been proposed. An arbitrary kernel can be plugged in to determine the similarity between two elements of the sequences, and the kernel value is determined as the normalized sum of the similarities along the backtracked longest common sequences (ALCS).

The authors of [12] use MKL to learn weights for different lengths k of k-mer string kernels for sequence classification in computational biology. They propose a MKL training strategy that is optimized for the weighted degree (WD) kernel, a specific string kernel. For combining different features, they sum over the components in a stacked feature vector. MKL has been successfully applied to image and video concept classification [13]. The use of MKL for feature selection in a visual classification task is proposed in [14]. A MKL formulation that learns feature weights together with the model, sampling features from a potentially infinite parameter space is used. Similarly, [15] use MKL for combining feature for video concept labeling. In [16] this approach is extended to learning per-sample feature weights. In [17], the authors propose an incremental MKL approach for visual concept classification. Based on a classifier trained on a standard training set, MKL is used to adapt the classifier to a specific scene. The authors of [18] propose adaptive MKL to adapt classifiers trained on web video to consumer video collections. Recently, the authors of [19] have proposed to use MKL for learning a combination of visual and audio features for concept classifiers in video. However, to the best of our knowledge, the application of MKL to sequence-based kernels in the multimedia domain has not been researched, including combining single-sample and sequence-based kernels.

II. Learning a Combination of Sequence-based Kernels

Multiple kernel learning (MKL) is an approach that considers a set of kernels potentially appropriate for their problems, and estimates both the parameters of the individual kernels as well as their relative weights during the training phase [20]. In particular, we discuss $L1$-norm MKL, which defines the kernel to be learned as a linearly weighted sum of different kernel functions. The authors of [14] discussed how this approach can not only be applied to combining different types of kernels, but also to combining a set of instances of kernels with different parameters, where the parameters can include features, parameters for feature extraction, kernel parameters, etc. The parameter space can thus be potentially infinite.

We can express any combination of sequence alignments based on different alignment types, alignment parameters or underlying features as a problem of learning weights of different sets of parameters of the kernels. To solve the MKL problem, we use the interleaved optimization method described in [21]. In the following, we discuss specific cases of applying the MKL approach to variation of alignment parameters and features.

A. Multiple alignment kernels

Sequence-based kernels combining multiple alignments can be rewritten as a combination of single alignment kernels. In this case, the features are fixed for all parameter sets, and we only vary the parameters controlling the alignment. For the TSGA, ASS and ALCS kernels the different alignments are weighted equally. For the DTAK kernel the weights are binary, selecting one alignment. The EMD kernel considers a weighted sum of all possible alignments, but as the weights are contained in the flow coefficients, the alignments can be weighted equally in the MKL formulation.

The MKL formulation allows the definition of variants of these multiple alignment kernels, which assign different weights to the alignments considered by the kernel. As the sum of the weights is normalized, this does not change the properties of the different multiple alignment kernels, as it results in multiplying the weight of an individual alignment with a scalar in the range $[0; 1]$. This approach enables better adjustment of multiple alignment kernels by learning the weights for the alignments. However, for kernels such as the ASS kernel that consider all possible complete and partial alignments between the input sequences, the formulation as MKL problem might be computationally very demanding, especially if the sequences to be processed are long.

B. Combining different features

For the feature vector of a single element in the sequence, combining different features using the proposed MKL formulation is equivalent to a weighted sum of the kernels evaluated on the different features in the vector. However, the latter would require that all features are sampled at the same time points and that the weights chosen for an alignment apply to all features, i.e., that an optimal alignment for one feature is also optimal for the others. Using the MKL formulation on the sequence-based kernels is much more flexible, as it can combine differently sampled features and selects the weights for their alignment separately. The proposed formulation allows not only varying the alignment parameters for each of the features, but also choosing a different combination of sequence-based kernels for each of the features.

For sequence-based kernels that are not guaranteed to be positive semidefinite under all conditions, summing the feature vectors may result in a higher fraction of negative eigenvalues of the kernel matrix, if some of the features used are not sufficiently discriminative (this problem is usually negligible when using a product to combine the kernel functions of the individual features). This may result in slow convergence during training.

III. Experimental Results

We have performed experiments on the TRECVID [1] 2011 Semantic Indexing (SIN) data set, using the 50 concepts of the light run. In order to apply sequence-based kernels, we have sampled more key frames based on visual activity than provided in the TRECVID master shot reference. From the key frames, we have extracted two MPEG-7 [22] descriptors globally (Color Layout and Edge Histogram) and bag of visual words features. For the latter, about 300 densely sampled image regions from 3 different scales are selected per key frame. A 128 dimensional SIFT descriptor (4×4 subregions, 8 directions for orientation histograms) is extracted for each of these regions without computation of a dominant orientation. We also extract MPEG-7 Color Layout features from these regions. We use codebooks with 100 codewords which leads to two 100 dimensional BoF features for each key frame. Besides global histograms of the entire key frames, we generated further versions where the key frames are split into 2×2, 1×3, 3×1, and 3×3 regions in horizontal and vertical direction, and a 100 bin feature histogram is extracted from each of the subregions.

For the MPEG-7 features we use the kernel proposed in [23], and for the bags of visual words we use the histogram intersection kernel [24]. For concepts that have a very high number of positive samples, the number of samples has been limited to the key frames of 1,000 shots (randomly sampled), and balanced with the same number of randomly selected negative samples. For solving the MKL problem we use the Shogun framework [25].

We parameterize 60 subkernels of an MKL problem. In a first experiment, the 60 parameter sets contain the 12 features described above (two global features, five SIFT and five Color Layout BoW histograms on different spatial

Concept	LCS, prod, $\theta = 0.03$	LCS, prod $\theta = 0.50$	LCS, prod $\theta = 0.90$	MKL max	MKL LCS $\theta = 0.10..0.90$	MKL LCS+max $\theta = 0.10..0.90$	fract. seq.
Adult	0.3616	0.2608	0.2305	**0.5712**	0.5041	0.5041	0.6607
Car	0.0055	0.0018	0.0003	0.0404	**0.0591**	0.0429	0.9442
Cheering	0.0009	0.0001	0.0001	**0.0071**	0.0001	0.0048	0.6987
Dancing	**0.0034**	0.0020	0.0000	0.0001	0.0029	0.0029	0.9870
Demonstration or Protest	**0.0629**	0.0268	0.0000	0.0001	0.0249	0.0249	0.9171
Doorway	0.0008	0.0004	0.0004	**0.0018**	0.0016	0.0016	0.8272
Explosion Fire	0.0001	0.0000	0.0000	**0.0009**	0.0002	0.0002	0.7960
Female Person	0.0011	0.0022	0.0024	0.0089	**0.0204**	**0.0204**	0.9235
Female-Human-Face-Closeup	0.0031	0.0008	0.0011	**0.0056**	**0.0056**	**0.0056**	0.8896
Flowers	0.0000	0.0000	0.0000	**0.0002**	**0.0002**	**0.0002**	0.9455
Hand	**0.0021**	0.0002	0.0000	0.0010	0.0010	0.0010	0.5382
Indoor	0.0215	0.0219	0.0152	**0.1491**	**0.1491**	**0.1491**	0.8703
Male Person	0.0159	0.0074	0.0089	**0.0335**	**0.0335**	**0.0335**	0.5011
Mountain	0.0032	0.0000	0.0000	**0.0414**	0.0381	0.0414	0.9524
News Studio	0.0050	0.0002	0.0001	**0.0105**	0.0104	**0.0105**	0.9482
Nighttime	**0.0055**	0.0002	0.0002	0.0050	0.0051	0.0050	0.9791
Old People	0.0102	0.0013	0.0008	**0.0373**	**0.0373**	**0.0373**	0.9860
Running	**0.0006**	0.0000	0.0000	0.0001	0.0001	0.0001	0.9619
Scene Text	0.0087	0.0044	0.0048	**0.0397**	**0.0397**	**0.0397**	0.8770
Singing	0.0002	0.0000	0.0000	**0.0006**	**0.0006**	**0.0006**	0.9811
Walking	0.0007	0.0003	0.0003	**0.0082**	0.0080	**0.0082**	0.8824
Walking Running	0.0245	0.0087	0.0089	0.0569	**0.0586**	0.0569	0.8402
mean	0.0244	0.0154	0.0125	0.0463	**0.0455**	0.0450	
median	0.0033	0.0006	0.0003	0.0077	0.0092	**0.0094**	

Table I

infAP of sequence-based kernels on combined features and MKL with subkernels using different features and thresholds (best value bold), and fraction of weight of sequence-based kernels for the combined single-sample/sequence-based kernel.

grids), each using the LCS kernel with 5 different values of the similarity threshold $\theta \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$. On the training data, we learn the models for each of the kernels as well as the relative weights of the subkernels. For comparison, we apply the LCS kernel to sequences, using a product of kernels for each of the individual features per key frame, with threshold $\theta \in \{0.03, 0.50, 0.90\}$. In a second experiment, we also use the same 12 features, but with a single-sample kernel using the pair of the best matching key frames to determine the similarity of two segments. We compare a MKL problem with subkernels for each of the features (MKL max in the table) and a MKL problem that uses both the five different parameterizations of LCS and the single-sample kernel (MKL LCS+max in the table).

Table I lists the inferred average precision (infAP[2]) for the 22 concepts of the light run of the TRECVID 2011 SIN data set, which are annotated in the ground truth and can thus be evaluated. For most of the concepts, the MKL approach provides at least as good results as the best parameterization of the LCS kernels, for some concepts the results improve significantly. The mean infAP over all the concepts doubles, and the median infAP increases by about 60%. The simple-sample kernel performs comparably well for many concepts, but is clearly outperformed by the sequence-based kernel for concepts involving dynamics such as Car, Dancing or Demonstration. However, overall the multiple sequence kernels and the combination of multiple sequence kernels with single-sample kernels yield very similar performance. The last column in Table I lists the fraction of kernel weights of the sequence subkernels for the combined MKL problem. It is apparent, that in cases where the sequence-based kernels outperform the single-sample ones, this fraction is rather high. However, the contrary is not true, i.e., a high fraction of weights for sequence-based kernels does not always indicate better performance of the sequence-based kernels.

IV. Conclusion and future work

Sequence-based kernels are promising representations for concept and event detection in video. In order to enable combining kernels working on heterogeneous types of features, using different ways of aligning sequences and parameterizations, and combine them with single-sample kernels, we propose the use of multiple kernel learning (MKL) with sequence-based kernels to video concept detection.

Experimental results on the TRECVID 2011 SIN data set show that using MKL on a set of kernels with different features and parameters significantly outperforms the same type of sequence-based kernel using an uniformly weighted combination of the same features, doubling the mean inferred average precision over all the concepts. For most the concepts, the MKL approach yields better results than the best parameterization of the same type of sequence-based kernel. While there are differences for individual concepts,

[2] Calculated using the revised TRECVID SIN script of Sept. 2012 addressing unequal number of positive and negative samples.

the MKL problem with only sequence-based kernels and that also including single-sample kernels yield the same mean and median results over the data set. Future work will address the combination of different types of sequence-based kernels as well as using this approach for fusing video and audio features.

Acknowledgment

The author would like to thank Robert Sorschag for providing some of the feature sets. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532, "TOSCA-MP".

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

[2] W. Bailer, "Sequence-based kernels for online concept detection in video," in *AIEMPro '11: Proceedings of the 4th international workshop on Automated information extraction in media production*, Dec. 2011, pp. 1–6.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, Dec. 2004.

[4] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. IEEE CVPR*, 2005.

[5] J. Choi, W. J. Jeon, and S.-C. Lee, "Spatio-temporal pyramid matching for sports videos," in *Proc. 1st ACM MIR*, 2008.

[6] D. Xu and S.-F. Chang, "Visual event recognition in news video using kernel methods with multi-level temporal alignment," in *IEEE CVPR*, 2007.

[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

[8] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," *CoRR*, vol. abs/cs/0610033, 2006.

[9] M.-C. Yeh and K.-T. Cheng, "A string matching approach for visual retrieval and classification," in *Proc. ACM MIR*, 2008.

[10] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video event classification using string kernels," *Multimedia Tools Appl.*, vol. 48, no. 1, pp. 69–87, 2010.

[11] W. Bailer, "A feature sequence kernel for video concept classification," in *Proceedings of 17th Multimedia Modeling Conference*, Taipei, TW, Jan. 2011.

[12] S. Sonnenburg, G. Rätsch, and C. Schäfer, "Learning interpretable SVMs for biological sequence classification," in *RECOMB 2005, LNBI 3500*, 2005, pp. 389–407.

[13] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[14] P. V. Gehler and S. Nowozin, "Let the kernel figure it out; principled learning of pre-processing for kernel classifiers," in *CVPR*, 2009, pp. 2836–2843.

[15] Y. Li, Y. Tian, L.-Y. Duan, J. Yang, T. Huang, and W. Gao, "Sequence multi-labeling: A unified video annotation scheme with spatial and temporal context," *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 814–828, 2010.

[16] J. Yang, Y. Li, Y. Tian, L.-Y. Duan, and W. Gao, "Per-sample multiple kernel approach for visual concept learning," *J. Image Video Process.*, vol. 2010, pp. 2:1–2:13, 2010.

[17] A. Kembhavi, B. Siddiquie, R. Miezianko, S. McCloskey, and L. S. Davis, "Incremental multiple kernel learning for object recognition," in *ICCV*, 2009, pp. 638–645.

[18] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *CVPR*, 2010, pp. 1959–1966.

[19] M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben, "Multimodal video concept detection via bag of auditory words and multiple kernel learning," in *Advances in Multimedia Modeling*, 2012, vol. 7131, pp. 40–50.

[20] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proc. ICML*, 2004.

[21] S. Sonnenburg, G. Raetsch, C. Schaefer, and B. Schoelkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, July 2006.

[22] "Information technology-multimedia content description interface: Part 3: Visual," ISO/IEC 15938-3, 2001.

[23] D. Djordjevic and E. Izquierdo, "Relevance feedback for image retrieval in structured multi-feature spaces," in *Proc. 2nd Intl. Conf. on Mobile Multimedia Comm.*, 2006.

[24] F. Odone, A. Barla, and A. Verri, "Building kernels from binary strings for image matching," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 169–180, 2005.

[25] S. Sonnenburg, G. Raetsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc, "The SHOGUN Machine Learning Toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, June 2010.