Unsupervised Visual Changepoint Detection using Maximum Mean Discrepancy

Michael Diu, Mehrdad Gangeh, and Mohamed S. Kamel

Centre for Pattern Analysis and Machine Intelligence,
Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada
{mdiu,mjabbarz,mkamel}@uwaterloo.ca
http://cpami.uwaterloo.ca

Abstract. The need to quantify similarity between two groups of objects is prevalent throughout the signal processing world. Traditionally, measures such as the Kullback-Leibler divergence are employed, but these may require expensive computations of covariance or integrals. Maximum mean discrepancy is a modern distance measure that is computationally simpler – involving the inner product between the difference in means of two groups' feature distributions – yet statistically powerful, because these distributions are mapped into a high-dimensional, nonlinear feature space using kernels, whereupon the means are estimated via the Parzen estimator. We apply this metric and leverage several powerful data representations from the supervised image classification world, such as bag-of-visual-words and sparse combinations of SIFT descriptors, to locate scene change points in videos with promising results.

Keywords: visual changepoint detection, unsupervised learning, maximum mean discrepancy, scene boundary detection, video indexing

1 Introduction

A vast literature has established that scene change detection algorithms have broad application in video indexing, analytics, summarization, and compression (see [20] for one representative survey). We introduce a novel method for detecting scene changes in videos, with several desirable properties — it is unsupervised, can work in an online or offline fashion, is not sensitive to thresholds or the genre of the video, allows for decimation of framerates and resolutions for high speed processing, and enables detection of different scenes, not just shot boundaries. It is tolerant to rotations, fast movement, and other non-semantic changes.

Work in this field has been underway for many years, under names such as video summarization, keyframe extraction, shot boundary detection, and changepoint detection. Inspired by the nomenclature of [3], such systems may be broadly decomposed and distinguished by the features used, their spatial support, the feature similarity metric employed, the region of temporal support chosen, and finally the boundary detection method.

Our system differs from others in two main ways. First, we adopt a more modern and powerful feature descriptor, the visual bag of words [4] using densely sampled scale-invariant feature transform (SIFT) keys [12] as the base words, which ensures robustness to noise, rapid motion, rotations, colour shifts, and global brightness/contrast changes. This approach has been shown to perform strongly in still-image scene recognition applications [11]. Secondly, we use a kernelized distance metric, the maximum mean discrepancy (MMD) [7] that is computationally simple, involving the inner product between the difference in means of the two distributions, yet statistically powerful, because these distributions are mapped into a high-dimensional, nonlinear feature space using kernels, whereupon the means are estimated via the Parzen estimator. The kernel representation allows us to efficiently use very high dimensional feature descriptors, by enabling computation of the MMD to occur in dissimilarity space and not using the original feature descriptors.

The MMD is computed over the frames of a video sequence in an overlapping sliding window fashion, successively forming 'current' and 'next' groups of frames. A standard peak finding routine is used on the MMD sequence to find local maxima, which are interpreted as scene change points. Previous work has demonstrated the usefulness of MMD and closely related variants in tasks such as speaker discrimination [8]; for segmenting musical notes [5], and EEG/ECG data [18]; and in matching small image patches for visual tracking [2].

To the best of our knowledge, neither of these two elements (the feature descriptor or similarity measure) have been previously proposed for scene boundary detection, although high-level video feature extraction [22] uses similar descriptors to summarize an entire video, obtaining its 'gist', without localizing the endpoints of individual scenes.

1.1 Previous Work

A review of the recent literature and surveys from the last decade ([20, 3, 9]) suggests continued research interest and activity in shot boundary and keyframe extraction techniques; fifty-seven groups and approaches participated in the annual NIST TRECVID [19] shot boundary detection competition held between 2001 and 2007. We suspect this popularity is in part because this application provides a nice testbed for new theoretical approaches in feature representations, dimensionality reduction techniques, distance measures, classifiers and clustering algorithms. Commercial applications of video indexing/summarization techniques include the keyframe summary that is available in the Google Youtube video service, and the multimedia indexing and search features for large corporate video collections in Autonomy Virage[1].

Based on the final TRECVID SBD workshop report [19], we may remark on some commonly observed characteristics of the top-performing shot boundary detection systems. They are based on local or global changes, as measured by

[1] `http://www.virage.com`

some similarity measure, of raw pixel feature vectors, colour histograms, or 2D-transformed versions of the frame (e.g. Fourier transform, wavelets), or of the frame's edges. One recent, representative work is in [13], which used wavelet features modeled using a generalized Gaussian distribution, and compared with the Kullback-Leibler divergence. The local maxima of the divergence in a fixed size window is marked as a shot boundary (*cluster* in their terminology) if it exceeds a threshold determined experimentally. However, a 'shot segment' or cluster derived using such an approach does not necessarily correspond to notable semantic changes in the video; rather, new shot segments may be formed due to camera or object motion.

A more general concern shared with other approaches, particularly parametric modeling approaches, is that there are a great number of parameters to be set. One way to handle this is to learn these parameters in a supervised manner; seven out of the top ten systems in the TRECVID 2005 competitions used SVMs [19], which require manually annotated groundtruth and may limit the system's ability to generalize to other video content genres.

Statistical models of scene change probability can also be built using supervised training data, and an *a posteriori* estimate of scene change probability computed, as in [17]. However as one might imagine, they are highly content-dependent (consider sports vs. nightly news) and also medium-dependent; that is, very different models may be seen in movies vs. surveillance video vs. mobile robotics.

Our approach differs from these works in major ways that have not been previously considered in the scene recognition field. We model the desired feature intent in a more direct fashion, borrowing from the scene recognition world, in order that the similarity measures may discriminate between different scenes more precisely while being less sensitive to intra-scene changes.

2 Methodology

2.1 Maximum Mean Discrepancy

Objects can be represented by either features or (dis)similarities. In a feature-based representation, a set of measurements (features) are computed based on expert knowledge domain. In (dis)similarity-based representations, objects are represented by their pairwise comparisons [14]. (Dis)similarity based representation can be computed on features or by comparing objects directly using a (dis)similarity measure [15]. There are cases, however, in which computing descriptive features to represent objects for a specific learning task is difficult or impossible due to insufficient knowledge on the domain [16]. To avoid this problem, objects can be represented in (dis)similarity space. In this approach, pairs of objects are compared by a (dis)similarity measure reflecting their mutual resemblance.

If we assume that objects whose dissimilarities are to be computed come from two different distributions, our ultimate goal is to find the distance between these

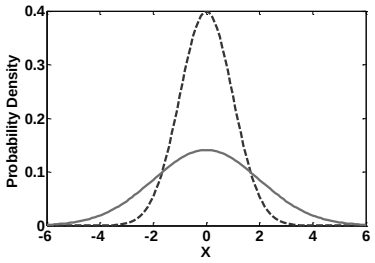

Fig. 1. Two Gaussian distributions with the same mean and different variances.

two distributions. One way to measure the distance or dissimilarity between two distributions is computing the distance between cluster means

$$\mathrm{d}(p,q) = \|\mathbf{E}(p) - \mathbf{E}(q)\|_2 \,, \tag{1}$$

where p and q are two distributions whose distances are to be computed, and $\mathbf{E}$ is the expectation function. The main drawback of using (1) is that it only takes into account the first order statistics of data samples taken from distributions. In case that the two distributions have the same mean (first order statistics), (1) cannot discriminate between them. An example is shown in Fig. 1. In this figure, p and q are two Gaussian distributions with the same mean but different variances and hence, using (1) by itself is not sufficient to discriminate them.

Given data samples $X = \{x_1, ..., x_n\}, X \sim p$ and $Y = \{y_1, ..., y_n\}, Y \sim q$, by using a mapping such as $X \xrightarrow{\varphi} \left[X\ X^2\right]$, we, effectively, bring data to a higher dimensional feature space. If we compute (1) in this augmented feature space, in fact, the second order statistics of the distributions are also taken into account. This, for example, can discriminate the two distributions given in Fig. 1. We use this basic concept to introduce maximum mean discrepancy (MMD) in reproducing kernel Hilbert space (RKHS). By using an appropriate kernel such as RBF kernel, this effectively maps data to a very high dimensional feature space and computing MMD in this space takes into account all statistics of the distributions up to infinity. This introduces a metric which is also called Hilbert Schmidt independent criterion (HSIC) that can discriminate two different distributions efficiently by mapping them to high dimensional feature spaces [6, 10].

Distance Based on Hilbert Schmidt Independent Criterion. Let's suppose that $X = \{x_i\}_1^n$ and $Y = \{y_i\}_1^m$ are data samples drawn independently and identically distributed (i.i.d.) from p and q, respectively. We define a feature mapping function φ such that $X \sim p, X \xrightarrow{\varphi} \varphi(X)$ and similarly $Y \sim q, Y \xrightarrow{\varphi} \varphi(Y)$ that maps data to a high dimensional feature space. By computing MMD in this

space, we compute a metric in RKHS with the following formulation

$$\begin{aligned}\mathrm{MMD}(\varphi, p, q) &= \|\mathbf{E}\{\varphi(p)\} - \mathbf{E}\{\varphi(q)\}\|_2 \\ &= \left[\mathbf{E}\{[\varphi(X) - \varphi(Y)]^\top [\varphi(X) - \varphi(Y)]\}\right]^{\frac{1}{2}} \\ &= \left[\mathbf{E}\{\varphi(X)^\top \varphi(X) - 2\varphi(X)^\top \varphi(Y) + \varphi(Y)^\top \varphi(Y)\}\right]^{\frac{1}{2}} \\ &= \left[\tfrac{1}{n^2}\textstyle\sum_{i,j} k(x_i, x_j) - \tfrac{2}{nm}\sum_{i,j} k(x_i, y_j) + \tfrac{1}{m^2}\sum_{i,j} k(y_i, y_j)\right]^{\frac{1}{2}} \quad (2)\end{aligned}$$

where $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$.

The last equation in (2) is the metric that measures the distance between two distributions using data samples which are drawn i.i.d. from these two distributions.

2.2 A Spatial Relationship-Preserving Scene Descriptor

We next turn our attention to the base feature vector used in the MMD computations when comparing one group of frames against another, that is, the scene descriptor that describes each frame. Given a greyscale input frame, SIFT keys are first computed on a dense grid spacing. The closest match to this key is found from a linear combination of visual 'words' in a dictionary, using a technique called locality-constrained linear coding (LLC) [21], and the coefficients of the linear combination used to increment each word's histogram bin by an amount proportional to their coefficients, thus forming the descriptor for the image. Rather than choosing a combination of words that minimizes ℓ_1 or ℓ_2 constraints, as in many sparse coding approaches, LLC regularizes the constraint equation using the distances between keys $\mathbf{x}_i$ and atoms $\mathbf{d}_i$, as shown in (3), where $\mathbf{B}$ is the dictionary of words, and $\mathbf{c}_i$ the coefficient vector ($\odot$ represents element-by-element multiplication).

$$\min_{\mathbf{C}} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{B}\mathbf{c}_i||^2 + \lambda ||\mathbf{d}_i \odot \mathbf{c}_i||^2 \qquad s.t.\ \mathbf{1}^T \mathbf{c}_i = 1, \forall i \quad (3)$$

The coefficients are constrained to be shift-invariant, which was found in [21] to have highest accuracy amongst the alternatives evaluated. The dictionary is built offline using the k-means unsupervised clustering technique with a dictionary size D. Larger dictionary sizes led to higher precision and recall, but are limited by increasing clustering time and memory requirements. Dictionaries from other videos may be used to enable scene change detection in online applications. Spatial relationships between words are utilized through the use of the spatial pyramid matching (SPM) technique, which, in a series of levels L, recursively subdivides the image by factors of four into regions, as illustrated in Fig. 2, and computes a LLC-based histogram within each region. These histograms are then concatenated together to form a final descriptor. We refer readers to [11] for further details.

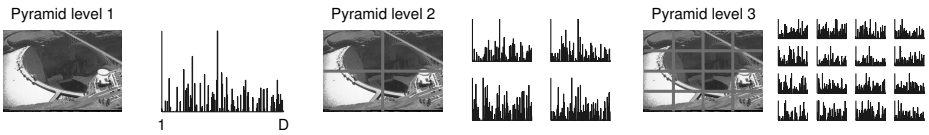

Fig. 2. An illustration of the spatial pyramid matching technique, showing a three-level pyramid of histograms. Each histogram bin represents one 'visual word'.

2.3 Scene Detection System

Finally, we discuss the integration of the MMD and the scene descriptor into a scene detection system, which is summarized in a block diagram, Fig. 3.

Aside from hard cuts from one scene to another, other common editing techniques that may confuse scene detectors (causing false negatives) are dissolves, fades, and wipes. Whereas many previous works have dealt with the problems of these special effects by essentially developing an independent detector for each special case [3], our scene descriptor deals with them in a simplified, unified, and elegant manner. Dissolves and wipes to a new scene, or fades to white/black, appear as prominent maxima in the MMD due to the significant change in the bag of words histogram. For accurate localization, we only need to ensure that the window size is longer than the longest expected fade. One benefit of the LLC-based feature descriptor is that dissolves and wipes involving two scenes may be directly modeled as a linear combination of words from the past and future scenes. This should, in theory, give us a smooth MMD peak rather than creating a noisy MMD signal by incorrectly allowing unrelated words to be matched to the image patch. Camera flashes are another kind of video content that can cause false positives, but may be dealt with by ensuring that there are no descriptors immediately on both sides of the large peak that have low mutual dissimilarity.

The histogram intersection kernel (HIK), (4), popularized in [11], is used to compare the descriptors of two images A and B, as it is parameterless and outperformed alternatives such as the RBF kernel in our tests.

$$K_{HIK}(A, B) = \sum_{j=1}^{r} \min\left(A_j, B_j\right) \tag{4}$$

The MMD is then computed as in (2) in a sliding window fashion for each frame t, with the $w/2$ frames prior to t forming group P, and the next $w/2$ frames forming group Q. The necessary SIFT keys and scene descriptors may also be computed in small batches which lends itself to online computation and reduced memory requirements. The common kernel computations between windows may be reused; only the kernel distance between each new frame and the $w-1$ other frames in its window need to be computed.

Finally, a standard peak finding approach is used to find local MMD maxima within the window. This falls in the class of adaptive thresholding methods, where a scene change is declared for a frame if it is a local maxima, and was preceded in time by a MMD value lower by at least δ.

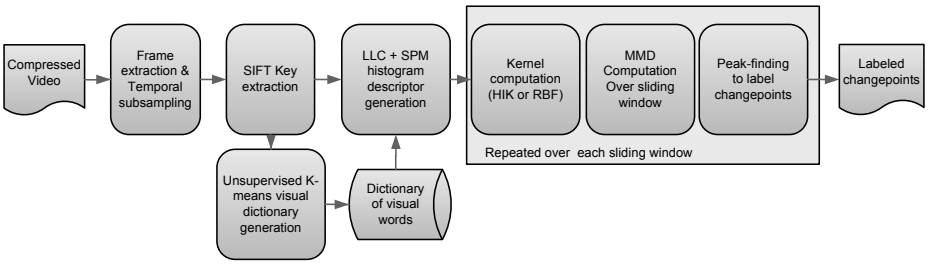

Fig. 3. Visual summary of our visual changepoint detection system.

3 Experimental Results

We demonstrate our system on a 30 minute documentary video used in the NIST TRECVID 2001 shot boundary detection competition, 'Challenge at Glen Canyon', featuring complex natural outdoor scenes, compression artifacts and noise, and a wealth of object and camera motion. SIFT keys are computed at every 8 pixels in both dimensions, with each SIFT key having a spatial support of 16×16 pixels. A visual dictionary of $D = 768$ visual words was learned by applying k-means clustering over 25% of the video. Using a three-level pyramid ($L = 3$), as recommended in [21], this yields a $768 + 4 \cdot 768 + 4^2 \cdot 768 = 16,128$ dimension feature descriptor for each image.

The top graph of Fig. 4 shows sample scene changes and the output MMD values. For video summarization and change detection purposes, we observe that it is unnecessary to locate the shot boundary with frame accuracy; a latency may be specified, which allows us to reduce computational complexity significantly by temporal subsampling. We hypothesize that the similarity operators used in systems based on the motion field, edges, or colour histograms would likely produce spurious false peaks due to the scene discontinuities, in contrast with our system. To investigate this we have decimated the framerate from 29.97 fps to 1 fps.

Results are shown in Table 3 using the well known precision and recall metrics, where Precision $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and Recall $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, using a set of 233 manually annotated groundtruth scene changes. TP, FP, and FN are the number of true positives, false positives and false negatives, respectively. The table entries are ordered by overall harmonic mean of the two, $F1 = 2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall})$. The first entry is the system as presented in Section 2. We set the window size to $w = 2$ experimentally with the aim of ensuring that we do not have any scenes shorter than this window. We examined the dynamic range R of the MMD over the video sequence and experimentally set the peak finding parameter $\delta = 0.05\,R$, but found the results are not too sensitive to this value.

We also wished to isolate the contribution of MMD on $F1$, from the contribution of the features. To test this, we treated the minima of the similarity kernel between successive frames without using MMD as scene changes, resulting in

entry 2 of Table 3 and the bottom graph of Fig. 4. We can see it is inferior in terms of precision and recall to the MMD solution.

In Section 1 we hypothesized that by computing the difference of group means in a remapped feature space, increased discrimination could be obtained. This is tested in entry 3 of Table 3, where the simple Euclidean distance (denoted by ℓ_2) is used as a similarity measure. One reason for its reduced performance can be visually seen in the middle plot of Fig. 4 – the dynamic range is much lower than with the kernel solutions, which makes it very sensitive to peak finding thresholds. Next, the fourth entry of Table 3 illustrates the improvement in performance gained from the LLC and SPM techniques compared to the base bag of words (BOW) approach. Finally, the last two entries of Table 3 test two representative schemes on the same data in order to illustrate their ability to adopt to different content genres and framerates. Entry 5 is a change-of-edge-intensity approach using fixed thresholds and local subregions compared with the sum of absolute differences (SAD) function, while entry 6 is a shot boundary detector tuned for a particular genre, sports videos [1]. Settings were left at their default values for both systems, and we observe that they do not perform well in the domain outside one that they were trained in, which suggests there is a role for our unsupervised, non-parametric robust scene detection scheme to fill.

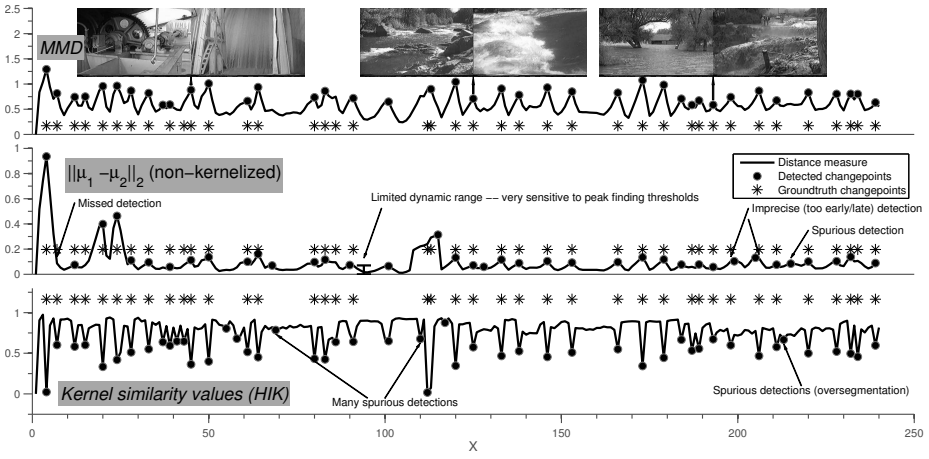

Fig. 4. Changepoints detected in the first four minutes of 'Challenge at Glen Canyon' film

4 Conclusions

We have presented the design and experimental results of a scene detection system that models the concept of 'scene' at a higher level of abstraction than

Table 1. Visual Changepoint Detection results on 'Challenge at Glen Canyon' video.

Approach	# Detections	Precision	Recall	$F1$
1) LLC + SPM + HIK + MMD	247	0.887	**0.940**	**0.913**
2) LLC + SPM + HIK	247	0.879	0.931	0.904
3) LLC + SPM : $\langle \underline{\mu_1} - \underline{\mu_2}, \underline{\mu_1} - \underline{\mu_2} \rangle_{\ell_2}$	255	0.839	0.919	0.877
4) BOW + HIK + MMD	245	0.853	0.897	0.875
5) Local edge-based SAD approach $\lvert\mu_1 - \mu_2\rvert > \tau$	730	0.315	**0.987**	0.478
6) Rank tracing approach [1]	158	0.544	0.369	0.440

previous works, using the visual bag of words approach and linear combinations of densely sampled SIFT keys. In doing so, elements of a video that do not represent a true scene change, such as object motion or contrast changes that other feature representations would be sensitive to, are treated as 'noise', reducing the false positive rate. The maximum mean discrepancy kernelized distance was then used with the histogram intersection kernel, with the aim to use nonlinear basis function expansion to increase separability of group means. Results showed higher dynamic range, precision, and recall compared to conventional methods.

Our system works well with varying framerates, enabling a CPU vs detection latency tradeoff, whereas most other scene change detectors are tuned with framerate-dependent thresholds on color or intensity features. We also illustrated how many special types of editing transitions can be handled naturally by the framework instead of requiring special cases and detectors. Three directions for future work are to expand the range of content tested, to use the MMD statistical significance tests in [7] to further filter out false positives, and to utilize the visual changepoint information and temporal segmentation information in the supervised problem of *place recognition* in videos; that is, to enforce temporal consistency in classification results by averaging out individual false positive classifications over a contiguous segment.

The contribution of this work, then, has been to demonstrate the application of the MMD to time series of visual data, and to 'marry' powerful feature descriptors from the world of computer vision with a problem based in the more traditional image processing world.

References

1. Abd-Almageed, W.: Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. In: 2008 15th IEEE International Conference on Image Processing. pp. 3200–3203. IEEE (2008)
2. Arif, O., Vela, P.: Robust density comparison for visual tracking. In: Proceedings of the British Machine Vision Conference. pp. 45.1–45.10 (2009)
3. Cotsaces, C., Nikolaidis, N., Pitas, I.: Video shot detection and condensed representation. a review. IEEE Signal Processing Magazine 23(2), 28–37 (2006)
4. Csurka, G., Dance, C.: Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, European Conf. on Computer Vision. pp. 1–22 (2004)

5. Desobry, F., Davy, M., Doncarli, C.: An online kernel change detection algorithm. IEEE Transactions on Signal Processing 53(8), 2961–2974 (2005)
6. Gretton, A., Borgwardt, K., Rasch, M.: A kernel method for the two-sample problem. Tech. rep., #157, Max Planck Institute for Biological Cybernetics (2008)
7. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A Kernel Method for the Two-Sample-Problem. In: Adv. in Neural Information Processing Systems 19. vol. 19, pp. 513–520. MIT Press (2006)
8. Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., Cappe, O.: A regularized kernel-based approach to unsupervised audio segmentation. In: IEEE Int'l Conf. on Acoustics Speech and Signal Processing. pp. 1665–1668. IEEE Computer Society (2009)
9. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A Survey on Visual Content-Based Video Indexing and Retrieval. IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews 41(6), 797–819 (2011)
10. Jegelka, S., Gretton, A.: Generalized clustering via kernel embeddings. In: Proc. of the 32nd German conf. on Advances in artificial intelligence. pp. 144–152 (2009)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Conf. on Computer Vision and Pattern Recognition. pp. 2169–2178 (2006)
12. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
13. Omidyeganeh, M., Ghaemmaghami, S., Shirmohammadi, S.: Video Keyframe Analysis Using a Segment-Based Statistical Metric in a Visually Sensitive Parametric Space. IEEE Transactions on Image Processing 20(10), 2730–2737 (2011)
14. Pekalska, E., Duin, R.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications. World Scientific, Singapore (2005)
15. Pekalska, E., Paclik, P., Duin, R.P.W.: A Generalized Kernel Approach to Dissimilarity-based Classification. Journal of Machine Learning Research 2(2), 175–211 (2002)
16. Pkalska, E., Duin, R.P.: Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters 23(8), 943–956 (Jun 2002)
17. Ranganathan, A.: PLISS: labeling places using online changepoint detection. Autonomous Robots 32(4), 351–368 (2012)
18. Sinn, M., Ghodsi, A., Keller, K.: Detecting Change-Points in Time Series by Maximum Mean Discrepancy of Ordinal Pattern Distributions. In: Conf. on Uncertainty in Artificial Intelligence (UAI). pp. 786–794 (2012)
19. Smeaton, A.F., Over, P., Doherty, A.R.: Video shot boundary detection: Seven years of TRECVid activity. Computer Vision and Image Understanding 114(4), 411–418 (2010)
20. Snoek, C., Worring, M.: A review on multimodal video indexing. In: Proc. IEEE Intl. Conf. on Multimedia and Expo. vol. 2, pp. 21–24. IEEE (2002)
21. Wang, J., Yang, J., Yu, K., Lv, F.: Locality-constrained linear coding for image classification. In: Conf. on Computer Vision and Pattern Recognition. pp. 3360–3367 (2010)
22. Yanagawa, A., Chang, S.: Columbia university's baseline detectors for 374 lscom semantic visual concepts. Tech. rep., #222-2006-8, Columbia University (2007)