IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 7, pp. 1251-1264, July 2014.
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6727470

Video tomographs and a base detector selection strategy for improving large-scale video concept detection

Panagiotis Sidiropoulos, *Member, IEEE*, Vasileios Mezaris, *Member, IEEE* and Ioannis Kompatsiaris, *Senior Member, IEEE*

***Abstract*—In this work we deal with the problem of video concept detection, for the purpose of using the concept detection results towards more effective concept-based video retrieval. The key novelties of this work are: 1) The use of spatio-temporal video slices (tomographs) in the same way that visual keyframes are typically used in video concept detection schemes. These spatio-temporal slices capture in a compact way motion patterns that are useful for detecting semantic concepts and are used for training a number of base detectors. The latter augment the set of keyframe-based base detectors that can be trained using different frame representations. 2) The introduction of a generic methodology, built upon a genetic algorithm, for controlling which subset of the available base detectors (consequently, which subset of the possible shot representations) should be combined for developing an optimal detector for each specific concept. This methodology is directly applicable to the learning of hundreds of diverse concepts, while diverging from the "one size fits all" approach that is typically used in problems of this size. The proposed techniques are evaluated on the datasets of the 2011 and 2012 Semantic Indexing Task of TRECVID, each comprising several hundred hours of heterogeneous video clips and ground-truth annotations for tens of concepts that exhibit significant variation in terms of generality, complexity, human participation. The experimental results manifest the merit of the proposed techniques.**

I. Introduction

The main goal of the video analysis community is the development of techniques that make possible the automatic understanding of the visual content and the semantic information conveyed by unconstrained video streams. By "unconstrained" we mean here videos that are not restricted to a specific known domain (e.g. soccer videos), and therefore can vary significantly both in their low-level visual properties and in their interpretation. Since for such unconstrained videos there is no restricted vocabulary that would be sufficient for describing their content and meaning, in the root of the efforts of the analysis community lies the development of methods for the fast and accurate detection of large numbers of diverse high-level video features, termed concepts. Concept detection, under this definition, means estimating for each concept a degree of confidence in the hypothesis that this concept is suitable for describing the contents of a given elementary piece of a video stream.

The multitude and diversity of the concepts that need to be detected, given the unconstrained nature of the content, creates significant challenges both in terms of effectiveness and computational efficiency. These challenges are highlighted, for instance, in the TRECVID Semantic Indexing task [1], which has been focusing on the development and benchmarking of systems that would be able to handle large amounts of video data and detect hundreds of semantic concepts efficiently (e.g. [2], [3]). As a result of this and other efforts towards solving the large-scale concept detection problem, several powerful techniques have emerged. For example, in order to exploit color information in addition to local image structure, the Opponent-SIFT and RGB-SIFT (or Color-SIFT) variations of the well-known SIFT descriptor [4] were proposed in [5]. In order to further reduce the computational cost of extracting such local descriptors, techniques such as Speeded Up Robust Features (SURF) [6] and DAISY [7] were introduced as fast SIFT approximations, whereas in parallel the use of corner detectors for interest point detection (e.g. Harris-Laplace [8]) has in many schemes been either replaced or complemented by dense sampling (i.e. the sampling of image patches on a regular dense grid). At the front of machine learning, which is key to finding the mappings between such low-level features and the high-level concepts that we want to detect, similar effectiveness and efficiency considerations have lead to interesting developments; for instance, chi-square kernels, that were originally considered to be optimal for use in support vector machines (SVM) [9], [10], are now often replaced by Histogram Intersection kernels [11] or even Linear SVMs for the sake of scalability.

Contrary to what is intuitively expected, in most of the developed schemes that aim to detect multiple concepts in large-scale video data, motion information is ignored and the detection is based exclusively on processing characteristic keyframes that are extracted at shot level (i.e. each video shot is represented by one or more keyframes). This is partially explained by the high computational cost associated with the extraction of most motion descriptors (since this extraction ne-

P. Sidiropoulos is with the Mullard Space Science Laboratory / University College London, UK. This work was carried out while he was with the Information Technologies Institute / Centre for Research and Technology Hellas, Greece.

V. Mezaris and I. Kompatsiaris are with the Information Technologies Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece, {psid, bmezaris, ikom}@iti.gr.

This work was supported by the European Commission under contract FP7-287911 LinkedTV.

cessitates, at the very minimum, the processing of a significant number of frames per shot), compared to the disproportionately low gains in accuracy that such descriptors introduce over the simpler keyframe-based approach. However, taking some form of motion information into account when trying to make sense of the video content is evidently desirable, and finding a way to do this with a computational cost comparable to that of processing a single keyframe represents a major challenge.

Moreover, in order to reach high accuracy levels, current approaches to concept detection typically describe each keyframe with numerous different low-level features, which are then used separately for building a number of base detectors that produce an abundance of intermediate detection results for each concept; these intermediate results are subsequently combined in order to generate the final concept detection output (e.g. [12], [13], [5]). While this strategy of generating an over-complete description of each keyframe, rather than using a single low-level descriptor alone, has been shown to boost detection accuracy, one can intuitively assume that not all possible descriptors and intermediate detection results are of equal importance (or, of any importance) for the detection of all possible concepts. Nevertheless, little attention has been paid so far to the automatic selection of an optimal subset of such intermediate results for each specific concept, which could introduce improvements in both the effectiveness and the computational efficiency of the concept detection process.

In this work we propose the use of video tomographs [14] (i.e. spatio-temporal slices with one axis in time and one in space) to represent video motion patterns. These tomographs are effectively and straightforwardly extracted from the video stream. Contrary to previous works on tomographs, we treat them as another form of keyframes and we apply to them the usual pipeline of feature extraction and transformation, for building a number of tomograph-based base detectors for each concept. As we will show, the processing of these tomographs is not computationally costlier than processing a single keyframe and, when used in combination with visual keyframes, can significantly enhance concept detection accuracy. Moreover, to take advantage of the diversity of the concepts that need to be detected, and of the possible redundancy that exists in a typical over-complete shot representation (comprising multiple visual descriptors extracted from keyframes and tomographs) with respect to detecting any single one of these concepts, we introduce a novel technique built upon a genetic algorithm that selects for each concept independently the respective optimal base detector subset, instead of invariably using all possible base detectors for all concepts. As will be shown, this again has benefits in both detection accuracy and computational cost.

The rest of the paper is organized as follows. Related work is reviewed in Section II. Visual tomographs and their proposed use for video concept detection are described in Section III, followed in Section IV by the introduction of the proposed base detector selection technique. Experimental results and comparisons on two large-scale datasets are presented in Section V and, finally, conclusions are drawn in Section VI.

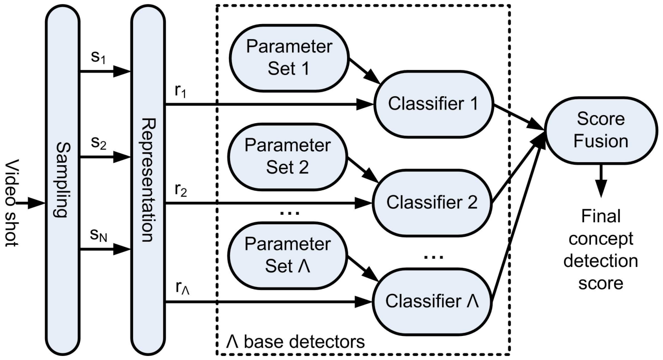


Fig. 1. The pipeline of a typical concept detection system. Initially the video stream is sampled (e.g. keyframes are extracted) using N different sampling strategies (labeled s_1, s_2,... s_N in this figure). Subsequently, Λ sets of features are extracted to represent the visual information samples (labeled r_1, r_2,...r_Λ). The set of features are used as inputs to base detectors that are trained offline. Finally, the base detector outputs are combined and an overall concept detection score is estimated.

II. Related work

The pipeline of a typical concept detection system is shown in Fig. 1. The video stream is initially sampled, for instance by selecting one or multiple keyframes per shot. Subsequently, each sample is represented using one or more types of features (e.g. SIFT [4], SURF [6]). These features form the input to a number of classifiers (i.e. base detectors), which typically are support vector machines. The parameter sets that control the employed classifiers are predefined, i.e. have been learned at the classifier training stage for each concept, using similar features extracted from training data. Finally, the base detector outputs are fused to estimate a final detection score for each concept.

From the above description it becomes apparent that large-scale concept detection systems encompass multiple video and image analysis techniques. These include techniques for interest point selection in keyframes [15] or video volumes [16], [17]; image descriptors such as Color-SIFT and Opponent-SIFT [18] or also spatio-temporal descriptors [19], [20]; vector modeling and quantization [21], [12]; code-book construction [22]; classifier selection and parametrization [11], etc.

However, little work has been done in the first component of the pipeline, i.e. the sampling of the video stream. The most common approach is to use a single keyframe per shot (e.g. [5], [2]), thus transforming the video concept detection task into image concept detection. This approach is fast and can straightforwardly take advantage of feature extraction techniques that were developed for still images, but it does not take into account motion, which may cause the visual content of a single shot to vary significantly. On the contrary, the authors of [12] propose using all video frames as keyframes. Whereas this enhances the accuracy, it also disproportionately increases the computational cost. A better balance between accuracy and efficiency can be achieved by using a limited number of keyframes per each shot, as in [3], in which the use of up to 10 keyframes per shot was proposed.

Techniques that employ only keyframes as video samples handle the video stream as a mere collection of photos (keyframes), failing to take advantage of the dynamic nature of video that makes it particularly expressive. Consequently, several techniques that involve also the motion modality have been introduced (e.g. [23], [24]). However, motion descriptor extraction is typically associated with high computational cost, and the gains in precision that are attained by introducing motion descriptors in the concept detection process are often not on par with the added computational complexity. In a recent review [19] a number of spatio-temporal interest point detectors and descriptors were examined both in terms of accuracy and computational cost and it was reported that the fastest among them could process 4.6 frames per second. In this case (assuming a mean shot duration of $5 - 7$ seconds and a frame rate of 25 frames/second), the motion descriptor extraction procedure would require more than half a minute per shot, a significantly longer time than that associated with the estimation of the most common image-based descriptors (e.g. SIFT, SURF) at keyframe level. Even if no $3D$ interest point detection is performed and dense sampling of the video volume is used instead (as proposed in [17]), the required computational time is still much higher than that associated with descriptor extraction at keyframe level. To alleviate this drawback in [23], [24] fast global motion descriptors were used (e.g. motion histograms) but these failed to enhance the detection accuracy. The authors of [24] also proposed the sampling of the video volume in spatio-temporal samples called short-term audio visual atoms (S-AVA); although the use of S-AVA instead of keyframes resulted in higher accuracy, the S-AVA estimation is based on a sophisticated and computationally demanding point tracking algorithm that has to be applied to all video frames. In general, even the fastest motion descriptors cannot compare with the fastest keyframe-based ones in terms of computational cost, often being practically inapplicable to large-scale video analysis problems.

Another part of the concept detection pipeline that has received little attention is the selection of the most appropriate set of features for each concept, rather than adopting an "one size fits all" strategy. For instance, approaches such as those discussed in the above paragraph can possibly increase the detection accuracy for motion-related concepts. However, large-scale video concept detection techniques should be able to handle multiple concepts which may or may not be related to motion (or to any other specific low-level visual property). Despite this need, in most of the relevant literature a process such as the one illustrated in Fig. 1 is executed invariably for all concepts of interest, despite their qualitative differences. More specifically, the complete base detectors set is employed for each and every concept and the base detectors scores are combined to estimate the overall output through averaging (e.g. [5], [3]) or linear combination using weights that are globally tuned for the complete set of concepts (e.g. [2]); in only a few approaches, these weights are selected independently for each concept through an off-line tuning process. For example, in [12] a brute-force search using cross-validation was proposed for selecting the weight values, which is computationally feasible only because the number of the employed base detectors in [12] is very limited. On the other hand, when a large number of base detectors are employed, weights may be tuned through gradient ascent [25] or regression analysis [13].

III. Employing video tomographs as additional shot samples

A. Use of tomographs

In this work we show how keyframe-based concept detection can be improved by augmenting the set of keyframes with a spatio-temporal type of image, the video tomograph. Video tomographs were introduced in [14] as spatio-temporal slices and have been used for optical flow estimation [26], camera motion classification [27] and video copy detection [28], [29]. A video tomograph is defined in [14] as a cross-section image (i.e. an image defined by the intersection between a plane and the video volume) which is additionally smoothed using a high-pass filter. The cross-section image is generated by fixing a 1-D line on the image plane and aggregating the video content falling on the corresponding line for all frames of the shot. In this work video tomographs are re-defined in a slightly different and somewhat more general way, and are used in a completely new way for a different application. A preliminary version of the proposed definition and use of tomographs was introduced by the authors in [30].

Video tomograph re-definition is based on the fact that the video volume is not continuous, but is formed by a finite set of frames. Consequently, a tomograph can be defined as a set of line segments, which are recursively estimated as intersections between lines and frames. More specifically, if f_i is the current frame, v_{i-1} the line defining the intersection in the previous frame, R_i the current tomograph rotation matrix and T_i the current tomograph translation vector then the $i-th$ line can be estimated as:

$$v_i = f_i \cap (R_i \cdot v_{i-1} + T_i) \quad (1)$$

If v_0 is the initial line segment and all R_i, T_i are known, then a tomograph image can be straightforwardly extracted. This tomograph definition encompasses the definition of [14], where the latter corresponds to setting $R_i = I_2, T_i = [0\ 0]^T\ \forall i$ in Eq. 1, I_2 being the two-dimensional identity matrix and superscript T denoting the transpose matrix. The advantage of the above definition is that complex motion patterns can be projected into meaningful images. For example, a tomograph could be formed by lines chosen so as to be always perpendicular to the camera motion direction, thus generating an image that captures the objects being followed by the camera. Such an approach would require knowledge of the camera motion, but this is not prohibitive since several methods exist for automatically detecting camera motion parameters from the video (e.g. [31], [32]), including methods that can be applied directly on the motion vectors encoded in the MPEG stream (e.g. [33]) and therefore having limited computational cost.

Putting aside the possibility of taking into account camera motion, the two simplest tomograph images are the centralized horizontal (CH-tomograph) and the centralized vertical (CV-tomograph) one. A CH-tomograph is constructed by aggregating into a 2D image the visual content of the horizontal line passing from the frame center, for all frames of a shot (i.e. $R_i = I_2, T_i = [0\ 0]^T\ \forall i$ and v_0 is the line $y = H/2$, where H is the frame height). A CV-tomograph is constructed in an analogous way, with the only difference being that v_0 is perpendicular to the x-axis, instead of parallel to it. Examples of CH- and CV-tomographs are shown in Fig. 2. In the left example the shot shows a man that crosses the camera field of view following direction from right to left. His motion is mapped into the CV-tomograph as a human silhouette, which regardless its slight deformation is human-understandable and can be processed using the same tools that are used for traditional images or keyframes. On the other hand, the main theme of the video shot of the right example is cloud movement. In this case the CH-tomograph achieves to express this motion into another human-understandable image, depicting primarily the moving clouds. We should stress here that, in the presence of arbitrary object or camera motion, the extracted tomographs are generally not expected to be human-understandable images; nevertheless, they do contain information that can help with concept detection, as our experiments will show.

For the purpose of concept detection in this work, both the CH- and CV-tomographs are processed in the same way as keyframes. More specifically, image patches are estimated, followed by descriptor extraction and vector quantization. The vocabulary (visual words) employed at this stage is constructed by clustering local descriptors extracted from the corresponding tomograph type (e.g. a random sample of CV-tomograph SIFT vectors are clustered in order to generate the vocabulary used for vector quantization of SIFT descriptors in CV-tomograph images). The resulting Bag-of-Words (BoW) feature vectors are the input to tomograph-based Support Vector Machine (SVM) classifiers (i.e. base detectors). These classifiers are also independently trained for each tomograph type, using annotated samples taken from tomographs of the corresponding type. As will be detailed in the experimental evaluation section, through this process we generate 24 tomograph-based base detectors for each concept (as opposed to only 12 in [30]), in addition to 12 similar keyframe-based ones (plus a 13-th keyframe-based one that uses simple color histograms rather than local image features). Finally, the output of all (i.e. up to 37) base detectors is fused following a simple late fusion scheme that does not discriminate between keyframe and tomograph-based detectors.

B. Computational concerns

Concerning the computational cost of introducing video tomographs in the concept detection process, it is straightforward that the processing time required for feature extraction from tomographs depends on the total number of pixels in each tomograph. Consequently, an estimation of the tomograph size can be used to compare the computational cost of tomograph-based classification with the computational cost of keyframe-based classification. Keyframe size is constant for a given video and can be adjusted during the decoding process. On the other hand, tomograph size is not constant, since it depends not only on frame size and frame ratio (that are typically constant) but also on the current shot duration. However, a rough estimation of the mean tomograph size is possible, at least for CH-tomographs and CV-tomographs. As a matter of fact, if W and H is the frame width and height, r is the frame rate and τ_s the duration of shot s then the total number of pixels for keyframe K, CH-tomograph K_H and CV-tomograph K_V would be:

$$pix(K) = WH, \quad pix(K_H) = \lfloor r\tau_s \rfloor W, \quad pix(K_V) = \lfloor r\tau_s \rfloor H \tag{2}$$

where $pix(.)$ denotes the number of pixels and $\lfloor . \rfloor$ the integer part of a real number. In the extensive TRECVID SIN 2012 dataset, the mean shot duration is 5.1 seconds. If typical values ($r = 25$, $\tau_s = 5.1$, $W = 352$, $H = 288$) are replaced in the above equations then the number of pixels of a CH- and a CV-tomograph together, compared to the number of pixels in a keyframe, would be:

$$(pix(K_V) + pix(K_H))/pix(K) \simeq 0.8 \tag{3}$$

Consequently, the descriptor extraction computational cost when using a pair of tomographs is similar to the cost of processing a single keyframe. Finally, it should be noted that the formation of the tomographs starting from the video stream is also very fast, particularly when simple CH- and CV-tomographs are used and therefore no processes such as camera motion estimation are required. In such a case, apart from decoding the video stream into frames, this only requires accessing and placing in a 2D matrix a small set of frame pixels (one fixed line segment per frame), as per Eq. (1).

IV. Selecting base detectors for building a concept classifier

A. Combining base detectors

The pipeline of Fig. 1 can be considered as a late fusion scheme that involves multiple base detectors executed independently, prior to combining their results. Late fusion represents the method of choice in state-of-the-art concept detection, as most related techniques use it either exclusively or in combination with some early fusion (e.g. [13], [12], [2], [5]). The use of multifarious information content as input to this fusion mechanism is guided by the need to be able to handle multiple concepts which may demonstrate significant diversity, for instance concepts that are either static (e.g. “forest”) or dynamic (e.g. “running”); rather specific (e.g. “George Bush”) or quite generic (e.g. “building”); human-based (e.g. “two people”), object-based (e.g. “motorcycle”) or background-based (e.g. “static background”); etc. Invariably fusing the output of all available base detectors for a given concept, as is typically the case in the literature, is based on the assumption that different base detectors can contribute to the accurate classification of concepts of a certain subset (e.g. detectors that use motion-based descriptors can contribute to detecting concepts that are related to motion), while they would not deteriorate the detection accuracy of all other concepts (e.g. it is assumed that having some motion-based detectors in the set of base detectors whose scores are fused would not adversely affect the final detection accuracy of static concepts). Thus, for each concept all base detectors scores

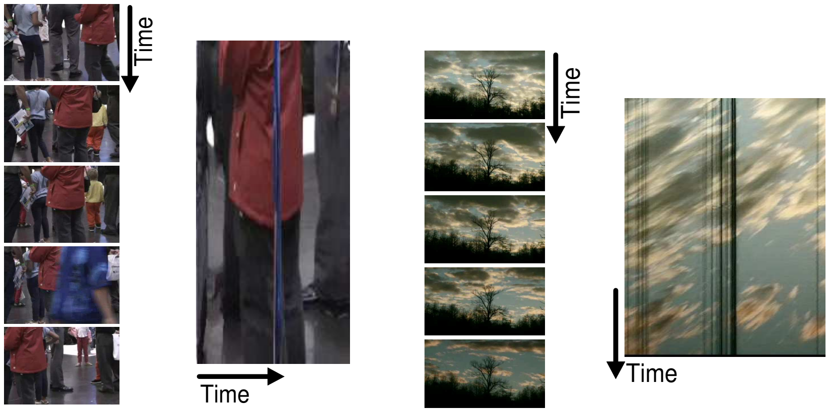


Fig. 2. Two tomograph examples, each one corresponding to a different type of tomograph image. The left tomograph is a CV-tomograph, while the right a CH-tomograph. Both of them are defined by the temporal ordering of lines that pass from the center of the frames. Five indicative frames of the shot from which each tomograph was generated are also shown to the left of the corresponding tomograph. The temporal order of the shown frames and the time-axis of the 2D tomograph images are denoted with arrows.

are combined to generate the final concept detection output. Consequently, assuming that the classification is performed at shot level, the associated complexity is proportional to $S \cdot D \cdot \Lambda$, where S is the total number of classified shots, D is the amount of concepts and Λ is the number of available base detectors.

The two most common late fusion strategies are averaging the confidence scores or calculating a linear combination of them with weights that are globally tuned for all concepts (e.g. [2]). The latter approach, while shown to enhance the concept detection accuracy, suffers from the "curse of dimensionality" that prohibits a brute-force tuning of the weights, especially since the number of base detectors is typically in the order of tens. Moreover, both late fusion approaches do not take into account the fact that not all of the employed detectors can truly contribute to the detection of all possible concepts. Therefore, computational time is unnecessarily consumed, since all base detectors are used in each concept, regardless if they increase the concept detection accuracy or not. Finally, as will be subsequently demonstrated, when the detection accuracy measure depends on the sorting of the results (as is the case with Average Precision (AP) or Extended Inferred Average Precision (XInfAP) [34]), averaging the confidence scores of two base detectors out of which only one achieves good concept detection accuracy may lead to the accuracy of the final detector (fused scores) being worse than the accuracy of the best out of the two original base detectors.

In order to clarify the last point, we model the confidence score distribution of a base detector that is trained to detect a specific concept as a mixture of two gaussian components $N(\mu_p, \sigma_p)$, $N(\mu_n, \sigma_n)$, each corresponding to true positive and true negative samples, respectively. The parameters that determine the classifier accuracy are $\mu_p, \sigma_p, \mu_n, \sigma_n$, as well as the concept's prior probability P_p. Since the Average Precision measure depends on the distance between the two gaussian components, the classifier is translation invariant. Thus, μ_n is arbitrarily selected to be equal to 0. Consequently, the performance of a detector C_3 that averages the scores of two base detectors C_1, C_2 is controlled by 7 parameters (parameters $\mu_p, \sigma_p, \sigma_n$ for each of C_1, C_2, and P_p, since the value of the latter depends only on the frequency of the concept in the dataset). In the subsequent analysis we examined 6 different values of P_p, $1\%, 5\%, 16.6\%, 33.3\%, 50\%$ and 66.6% in order to model concepts that are rare as well as concepts that can be found very often in a video stream. For each P_p value we estimate the probability of the following hypothesis being true: "$P(AP(C_3) > max(AP(C_1), AP(C_2)))$" (this probability we denote as P_0 in the sequel). We estimate the value of P_0 as a function of $AP(C_1)$ and $AP(C_2)$ (i.e. the average precision of the two base classifiers being combined), and we compare it with 0.5, since $P_0 > 0.5$ signifies that the case of detector C_3 performing better than the best detector among C_1, C_2 is more probable than the opposite one.

Plotting P_0 exclusively as a function of $AP(C_1)$ and $AP(C_2)$, for a chosen value of P_p, so as to reveal the relationship between P_0 and the AP scores of the base classifiers, is achieved through the following procedure. First, we randomly selected a large number of possible σ_p, σ_n value pairs (10^6 different pairs in our simulations). Then, for a given value of μ_p, we estimated the AP score of a base detector C_1 (or C_2) for each possible σ_p, σ_n value pair, and the average of these AP scores we considered as the AP score for the chosen value of μ_p (independently of σ_p, σ_n). Repeating this simulation for different values of μ_p, we found that this AP score is monotonically increasing with μ_p, and we built a lookup table allowing us to estimate the expected μ_p value that corresponds to a chosen AP score for C_1 (or C_2). Subsequently, P_0 can be plotted as a function of $AP(C_1)$ and $AP(C_2)$ as follows: (a) for each pair of base detector AP scores, the corresponding pair of μ_p values are found (using the lookup table), (b) a random pair of σ_p, σ_n values is selected, (c) using these parameter values the AP performance of the detector C_3 is estimated and compared with the maximum value of C_1 and C_2, and d) this process is repeated, similarly to when estimating the AP of a base detector above, for multiple pairs of σ_p, σ_n values. Following this, the value of P_0 is retrieved as the relative frequency of $AP(C_3)$ being greater than $max(AP(C_1), AP(C_2))$.

The results of the above simulation are demonstrated in Fig. 3. In this figure, starting from the top-left corner, the (i, j) block shows the value of P_0 when C_1 and C_2 base detectors have AP equal to $0.15 + 0.05i$ and $0.15 + 0.05j$, respectively. The grey level of each block represents the P_0 value, with white standing for $P_0 = 1$ and black for $P_0 = 0$. Moreover, a dot is drawn in the center of block (i, j) if $P_0(i, j) > 0.5$. It should be noted that we have chosen this method of visualization instead of reporting the exact P_0 scores, since the process we described above and used for getting these results is an approximate simulation for extracting mostly qualitative conclusions, rather than an exact mathematical model for the accurate quantitative analysis of AP. Fig. 3 demonstrates that the common assumption that the use of more classifiers always increases the accuracy of their combination is not to be taken as a general rule. On the contrary, if the detection performance already reached by a base detector is relatively high, combining additional base detectors with it or not needs to be thoroughly examined in order to make sure that the accuracy of the combination will not deteriorate, compared to using the single well-performing detector alone. More importantly, if an extensive set of base detectors is available for multiple concepts, a procedure for selecting for each concept the optimal subset of base detectors to be combined is needed.

Moreover, Fig. 3 manifests that P_0 exhibits a different behavior for concepts having different prior probabilities. Specifically, a significant difference between rare and frequent concepts is that in the first case accurate base detectors "dominate" over all other base detectors, since their combination enhances the performance even when the other base detector exhibits a low AP (e.g. for $P_p = 1\%$ or $P_p = 5\%$). On the contrary, when P_p increases, then the poor base detectors "dominate" the combination, since the performance remains low even when the other base detector of the combination exhibits a high AP. This point explains the disagreement in conclusions between participants in the TRECVID competition (that includes many rare concepts), who have found employing multiple base detectors and late fusion advantageous, and other works that, following the analysis of concept detection results for mostly frequent concepts, rejected such a late fusion approach (e.g. [24], [35]).

B. Base detector selection for concept detection

Motivated by the above analysis, we have developed a base detector selection procedure that selects for each concept the optimal subset of the available base detectors. More specifically, the introduced scheme builds upon a genetic algorithm, by post-processing the outcome of two different variations of the latter. We should stress that this algorithm is executed off-line, during training. The result of this technique is the selection for each concept of an optimal subset of base detectors and the exclusion of all other base detectors from the corresponding detection scheme. Thus, at run-time, the computational complexity of concept detection is reduced from $O(S \cdot D \cdot \Lambda)$ (Section IV-A) to $O(S \sum_{i=1}^{D} M_i)$, where M_i is the number of base detectors employed for the detection of the i-th concept.

The genetic algorithm that plays a central role in this approach is summarized in Alg. 1. In the following, we use the operator $\#$ to represent set cardinality, and the $\circ$ operator to represent the vector Hadamard product (i.e. the result of element-wise multiplication). Moreover, we denote L the ordered set of base detectors, L_i a subset of this set, p_i the AP achieved by using the average of the base detectors score that belong to L_i and v_i the participation vector of subset L_i. As participation vector of a subset L_i we refer to a binary vector of length $\#L$, whose j-th element is equal to 1 if and only if the j-th element of L belongs to L_i.

Algorithm 1 Genetic algorithm for selecting a subset of best-performing (or worst-performing) base detectors.

Notation: c is the current concept, m the mutation rate of the genetic algorithm, N is the initial population size, R the (fixed) number of generations and k is the number of parent chromosomes that breed the next generation population.

1: Initially, from set L, N random subsets (chromosomes) $L_1, L_2, ..., L_N$ are used to form the initial population. Their corresponding participation vectors $v_1, v_2, ..., v_N$, as well as the corresponding performance estimations $p_1, p_2, ..., p_N$ are computed. The current generation index r is set to 1.

2: The k chromosomes that achieved the best (or worst) performance ($k < N$) "survive", while all the other chromosomes are discarded.

3: Uniform crossover is used to combine the k parent chromosomes of the current generation in $k(k-1)/2$ pairs to breed two new chromosomes each, thus leading to a new population of $k(k-1)$ members. More specifically, from two parent chromosomes L_i and L_j the children chromosomes will have participation vectors $v_i \circ v_j + Y \circ (1 - v_i \circ v_j)$ and $v_i \circ v_j + (1 - Y) \circ (1 - v_i \circ v_j)$, where Y is a random binary vector of dimension $\#L$, $\sum Y = \#L/2$, and "1" denotes a vector of ones. Thus, each child chromosome inherits some of its genes from both its parents.

4: Once the new population is constructed mutation is employed to randomly modify a gene subset. More specifically, out of the $k(k-1)(\#L)$ genes of the population, $mk(k-1)(\#L)$ of them randomly mutate (i.e. the corresponding participation vector elements change value from 0 to 1 or from 1 to 0).

5: The chromosomes that match the $k(k-1)$ participation vectors formed after the end of step 4 are retrieved and the corresponding performance is estimated.

6: If $r = R$ then the chromosome L_0 that achieved the maximum performance is returned as the optimal configuration. Moreover, the participation vectors $v_1, v_2, ..., v_T$ of the chromosomes that achieved the top-T performance values are retrieved. Otherwise, $r = r + 1$ and the algorithm continues from step 2.

The aforementioned genetic algorithm is executed independently twice under the proposed approach. In the first execution the goal is to identify the configurations (i.e. the base detector subsets) that achieve the best performance, while in the second one the configurations that exhibit the worst performance. These are used for robustly selecting the best base

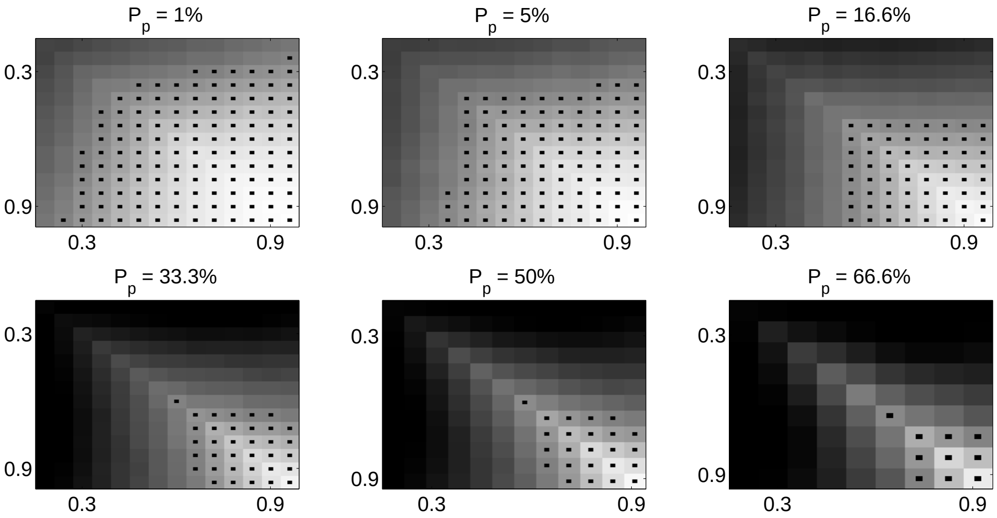


Fig. 3. The probability of a classifier that averages two base detectors to exhibit higher AP performance than the best-performing one of these base detectors, as a function of the AP of the two base detectors. The two axis correspond to the AP performance of each base detector. More specifically, starting from the top-left corner, the (i, j) block corresponds to a pair of detectors, having AP equal to $0.15 + 0.05i$ and $0.15 + 0.05j$, respectively. The grey level of each block represents the value of probability P_0, with white standing for 1 and black for 0. Moreover, a black dot in the center of a block (i, j) denotes that $P_0(i, j) > 0.5$.

detectors for the given concept: a base detector is ultimately selected if it is frequently included in the best-performing configurations and at the same time is not frequently included in the worst-performing ones. For realizing this selection, we introduce a "base detector quality" measure Q_c, in relation with concept c,

$$Q_c = (P_T - N_T)/T \tag{4}$$

where T is the number of configurations that are retrieved in each genetic algorithm execution, and P_T, N_T is the number of times that the base detector was included in the configurations that achieved the $top - T$ and the $bottom - T$ performance, respectively. The base detectors for concept c are ranked using Eq. (4), and the M highest-ranked ones are selected. The above described process is summarized in Alg. 2.

It should be noted that fixing the number of base detectors M_i that are selected for each concept to $M_i = M$, as we do in step 4 of Alg. 2, is a choice we make for simplifying the experimental evaluation and comparison of the proposed approach. In practice, one could also search for the optimal number of base detectors separately for each concept, which may further improve the results.

Algorithm 2 The proposed base detector selection algorithm.

1: For each concept the genetic algorithm (Alg. 1) is executed. In each generation of the algorithm the configurations that achieved the best performance "survive". The output is the $top - T$ configurations that achieved the best performance for each concept and the mean base detector number M (i.e. the average of the number of base detectors in the $top - 1$ configurations across all concepts, rounded to the nearest integer).

2: For each concept the genetic algorithm (Alg. 1) is executed again. This time, in each generation of the algorithm the configurations that achieved the worst performance "survive". The output is the $bottom - T$ configurations that achieved the worst performance for each concept.

3: For each concept and each base detector, the number of times that this base detector was included in the $top-T$ and the $bottom-T$ configurations (i.e. P_T and N_T, respectively) is estimated, and the base detector quality measure Q_c is calculated according to Eq. 4.

4: For each concept the M base detectors with the highest Q_c comprise the base detector subset (i.e. configuration) that will be employed in the concept classifier, while all other base detectors are discarded.

V. Experimental evaluation

A. Datasets and experimental setup

Our experimental setup is based on the 2011 and 2012 TRECVID SIN Tasks [36], [37]. As already mentioned, the total number of concepts that were defined in these tasks is 346. However, the corresponding evaluations were carried out in a subset of 50 and 46 concepts, respectively, for which ground-truth annotation of the test sets is available. The latter concepts, which are the ones used in this work, are shown in Table I. In this table we additionally mark with a "*" the concepts that are either directly related with motion (e.g. "throwing", "walking-running", "skating") or correspond to objects that are very likely to be filmed while they are in motion (e.g. "skier", "car", "boat-ship").

For training, validation and testing our concept detectors we have used the TRECVID SIN Task datasets as follows: For the experiments on the 2011 dataset our training set consisted of 11644 videos (lasting approximately 400 hours and including 250000 shots), the validation set of 4216 videos (approx. 100 hours; 70000 shots) and the test set of 4000 videos (approx.

TABLE I
THE 50 AND 46 CONCEPTS EVALUATED IN THE 2011 AND 2012 TRECVID SIN DATASET, RESPECTIVELY. MOTION-RELATED CONCEPTS ARE MARKED WITH A "*".

Dataset	Concepts
TRECVID 2011	Adult, Anchor-person, Beach, Car*, Charts, Cheering*, Dancing*, Demonstration*, Doorway, Explosion/Fire, Face, Female Person, Female Face Close-up, Flowers, Hand, Indoor, Male Person, Mountain, News Studio, Nighttime, Old People, Overlaid Text, People Marching*, Reporters, Running*, Scene Text, Singing, Sitting Down*, Sky, Sports*, Streets, Two People, Walking*, Walking/Running*, Door Opening*, Event*, Female Human Face, Flags, Head & Shoulder, Male Human Face, News, Quadruped, Skating*, Speaking, Speaking To Camera, Studio With Anchor-person, Table, Text, Traffic*, Urban Scenes
TRECVID 2012	Airplane*, Airplane Flying*, Basketball*, Bicycling*, Boat-Ship*, Boy, Bridges, Chair, Computers, Female Person, Girl, Government leader, Greeting*, Highway*, Instrumental Musician, Kitchen, Landscape, Male Person, Meeting, Motorcycle*, Nighttime, Office, Press Conference, Roadway Junction, Scene Text, Singing, Sitting Down*, Stadium, Teenagers, Throwing*, Walking-Running*, Apartments, Baby, Civilian Person, Clearing, Fields, Forest, George Bush, Glasses, Hill, Lakes, Man Wearing A Suit, Military Airplane*, Ocean, Skier*, Soldiers*

100 hours; 65000 shots). For the experiments on the 2012 dataset, the training set consisted of 19860 videos (approx. 600 hours; 400000 shots), the validation set of 4163 videos (approx. 100 hours; 73000 shots), while the test set of 4100 videos (approx. 100 hours; 72000 shots).

For evaluating the trained concept detectors we followed the methodology used in TRECVID. That is, for each concept separately, the top 2000 shots sorted by detection score in descending order are returned and are evaluated against partial manually-generated ground-truth annotations. The evaluation measure is the Extended Mean Inferred Average Precision (XInfAP) [34], which has been proposed for the purpose of approximating Average Precision (AP) when the dataset is not fully annotated and therefore AP cannot be directly calculated.

The proposed concept detection approach is implemented, for the purpose of experimental evaluation, according to the pipeline of Fig. 1, employing state-of-the-art features and parameters for the pipeline components that are not explicitly discussed in the preceding sections. Specifically, each video shot is represented by either one or more keyframes or also a pair of video tomographs (CV-tomograph, CH-tomograph). Subsequently, an interest point detector is employed to select the image points at which descriptors will be extracted. We used two such detectors; the first selects interest points through dense sampling, i.e. in fixed distances in a 2D image grid, while the second one is a Harris-Laplace corner detector [8]. At each of the resulting interest point locations, low-level visual descriptors are extracted (SIFT, RGB-SIFT and Opponent-SIFT), following the conclusions drawn in [5] for video concept detection tasks. Subsequently, the low-level descriptors are assigned to visual words, using two vocabularies that were created off-line through k-means clustering and hard- or soft-assignment, respectively, according to [38]. In all cases (i.e. regardless of which one of the above video sampling strategies, descriptors etc. is used) a pyramidal 3×1 decomposition scheme employing 3 equally-sized horizontal bands of the image, as proposed in [39], was used on every keyframe or tomograph, generating 3 different BoWs corresponding to the three image bands and a fourth BoW for the entire image. The number of words for each BoW was set to 1000 and the four BoWs coming from the adopted pyramidal decomposition were concatenated to a 4000-element BoW vector. One such vector is calculated separately for each combination of representation (i.e. each keyframe or tomograph), interest point detector, descriptor and assignment method and is used as input to the corresponding SVM base classifiers for a given concept (base detectors). As a result, 36 base detectors are built for each concept separately, based on local image features. An additional base detector that uses a global visual descriptor (HSV histogram) is also employed. The 37 base detectors are outlined in Table II.

For the base detectors, linear SVM classifiers are employed instead of the kernel SVMs that are typically used in such tasks. By this choice, the required computational time for a single SVM classifier (corresponding to a single concept) fell from 6 seconds per image (that was required in our earlier experiments with kernel SVMs) to 0.03 seconds. This is an implementation choice that is in line with our overall goal of developing a computationally efficient solution to concept detection. All base detectors were trained off-line, using the corresponding training sets. The output of each of the trained base detectors is an intermediate confidence score. The overall confidence score for each concept is estimated as the harmonic mean value of the intermediate confidence scores of all (i.e. up to 37) base detectors for this concept. Using the harmonic mean was shown in practice to produce slightly better results than the arithmetic mean, which was used for simplicity in our simulations reported in Section IV-A.

B. Tomographs versus keyframes

In this first experiment we investigate the impact of using the video tomographs introduced in Section III in replacement of traditional keyframes. More specifically, we compare the combination of the 13 keyframe-based detectors of Table II against the combination of 12 tomograph-based detectors of the same Table that use dense sampling for interest point detection (which, as we show later on, are the best among our 24 tomograph-based detectors). The results show that out of the 96 different concepts in the two datasets that we use, tomographs outperform keyframes for only 8 concepts (all of them motion-related: basketball, throwing, skier, demonstration, people marching, running, walking/running, skating). This result is consistent with our intuition that tomographs convey useful motion information that is, however, complementary rather than alternate to the non-motion information

TABLE II
COMPLETE SET OF BASE DETECTORS USED IN OUR EXPERIMENTS.

Video Sampling	**Possible representations and resulting base detectors**
Keyframes	12 local-image-feature-based Base Detectors: 3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) × 2 point detectors (Dense, Harris-Laplace) × 2 BoW strategies (soft-, hard-assignment)
	1 global-image-feature-based base detector (HSV color histograms)
Tomographs	24 tomograph-based base detectors: 2 types of video tomographs (CH-tomograph, CV-tomograph) × 3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) ×2 point detectors (Dense, Harris-Laplace) × 2 BoW strategies (soft-, hard-assignment)

conveyed by keyframes. Based on this result and in accordance with the emphasis that we put in all previous sections on how to combine tomographs and keyframes, we will proceed in the sequel with experimentally evaluating such combinations.

C. Combining tomographs and keyframes

In this series of experiments we investigate the impact of combining the video tomographs introduced in Section III with traditional keyframes. The following approaches are compared: i) $A13$ is the baseline, using 1 keyframe per shot and combining the 13 keyframe-based base detectors of Table II, ii) $A25$ combines the 13 base detectors of $A13$ with 12 additional base detectors that employ tomographs as video samples and use dense sampling for interest point detection, iii) $A37$ combines all 37 base detectors of Table II, iv) $A39$ employs the same base detectors as $A13$, but, as proposed in [3], in this case 3 keyframes per shot are used instead of one and therefore the 13 base detectors of $A13$ have to be evaluated 3 times each for a single shot (thus, 39 base detector scores are produced per shot). Our choice to compare tomographs with the multiple keyframe approach of [3] is based on the fact that any other existing solution (e.g. traditional motion descriptors) would feature significantly higher computational cost, making the application of it on the extensive TRECVID datasets difficult and in any case depriving such a comparison of any practical value.

The experimental results are summarized in the top block of Table III (MXInfAP, i.e. mean XInfAP across all concepts). Comparing $A25$ with $A13$, the MXInfAP increases in the 2011 dataset from 0.2187 to 0.2659 and in the 2012 dataset from 0.1394 to 0.1557. These improvements represent a 21.6% and 11.7% accuracy boost, respectively, and manifest that although the tomographs are not potential replacements of traditional keyframes, they provide additional information that the latter do not capture. Moreover, as can be seen in Table III, it is the dynamic (motion-based) concepts that benefit the most from the introduction of tomographs: if only the motion-related concepts are taken into account, then the accuracy boost caused by introducing tomographs is 38.9% and 47.1% for the two testsets, respectively.

On the other hand, the reported results show that the detection of interest points by means of Harris-Laplace in tomograph images generates noisy representations that lower the overall performance. This is manifested by the significantly lower performance that $A37$ exhibits compared to $A25$ (MXInfAP of 0.2113 versus 0.2659 and 0.0837 versus 0.1557 in the 2011 and 2012 datasets, respectively). This is intuitively supported by the fact that tomographs are by default very noisy images, thus selecting image corners as interest points in such images does not necessarily help. However, as we show in the next subsection, if our concept detection problem focuses on motion-related concepts and base detector selection can be performed for each concept, then base detectors of this type can also contribute to more accurate results.

We also compare $A25$ with $A39$, which performs a denser video sampling in the temporal direction [3], thus using a larger number of keyframes for each shot. While $A39$ employs 13 distinct base detectors (the same used in $A13$), the use of 3 keyframes per shot results in each base detector being evaluated 3 times for each shot and therefore needs computational time similar to that of $A37$ (we should note that "computational time" here and throughout the rest of the Experimental Evaluation section refers to the time needed for obtaining a classification result for a given non-classified shot). Despite the additional keyframes, the performance of $A39$ is significantly lower compared to that of $A25$, while the latter is also less computationally demanding (Table III). This shows that tomographs should be preferred over introducing additional keyframes in a keyframe-based concept detection scheme.

Finally, in Figs. 4 and 5 we show results per concept (XInfAP) for the $A13$, $A25$ and $A37$ experiments. Although many of the considered concepts are not intuitively expected to be strongly correlated with any type of motion (e.g. "landscape", "fields", "computers"), we can see that when combining keyframe-based and certain tomograph-based base detectors ($A25$), XInfAP increased for more than 80% of the concepts.

D. Selecting base detectors for building a concept classifier

In this series of experiments we investigate the impact of using a base detector selection strategy for choosing a different subset of base detectors for each concept that we want to detect. For this we apply the technique introduced in Section IV and evaluate it against the results of the previous set of experiments (where no such selection strategy was used) as well as against variations of the technique of Section IV and related literature works [13], [40], [25]. More specifically, starting from the 25 base detectors used in $A25$, the following approaches are compared: i) $G25a$, a variation of the proposed technique in which only the first step of Alg. 1 is executed and the chromosome from the initial population whose performance is maximum at the end of that step is returned as the chosen configuration (i.e. set of base detectors), ii) $G25b$,

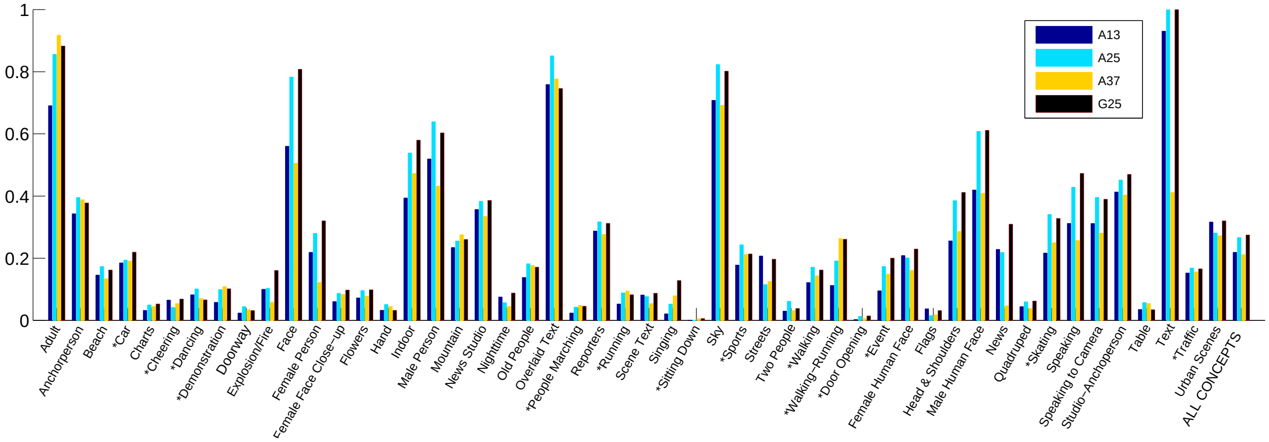


Fig. 4. XIinfAP for the concepts of the 2011 dataset. The compared techniques are $A13$, $A25$, $A37$ and $G25$.

TABLE III
EXPERIMENTAL RESULTS AND COMPARISONS. AVERAGE BOOST IS THE RELATIVE PERFORMANCE INCREASE COMPARED TO $A13$, WHILE # BASE DETECTORS IS THE AVERAGE NUMBER OF BASE DETECTORS PER CONCEPT AND SHOT THAT NEED TO BE EVALUATED AT RUN TIME.

	TRECVID 2011 dataset					TRECVID 2012 dataset				
	All concepts		Motion-related concepts only			All concepts		Motion-related concepts only		
Method	MX-InfAP	Average Boost	MX-InfAP	Average Boost	# Base Detectors	MX-InfAP	Average Boost	MX-InfAP	Average Boost	# Base Detectors
Tomographs and keyframes vs keframes only (experiments of Section V-C)										
$A13$ (baseline)	0.2187	0%	0.0959	0%	**13**	0.1394	0%	0.0797	0%	**13**
$A25$	**0.2659**	**21.6%**	**0.1332**	**38.9%**	25	**0.1557**	**11.7%**	**0.1172**	**47.1%**	25
$A37$	0.2113	-3.4%	0.1247	30%	37	0.0837	-39.7%	0.0845	6%	37
$A39$	0.2417	10.5%	0.1126	17.4%	39	0.1471	5.5%	0.0911	14.3%	39
Base detectors selection among 25 base detectors (experiments of Section V-D)										
$G25a$	0.2536	16%	0.1291	34.6%	10.32	0.1571	12.7%	0.1177	47.7%	10.11
$G25b$	0.2701	23.5%	0.1339	39.6%	10.02	0.1651	18.4%	0.1238	55.3%	8.89
$G25$	**0.2744**	**25.5%**	**0.1346**	**40.4%**	**10**	**0.1783**	**27.9%**	**0.1299**	**63%**	9
$R25$ [13]	0.2247	2.7%	0.1197	24.8%	25	0.1422	2%	0.1023	28.4%	25
$ABC25$ [40]	0.2617	19.7%	0.1282	33.7%	10.52	0.1697	21.7%	0.1282	60.9%	**8.87**
$GA25$ [25]	0.2585	18.2%	0.1301	35.7%	25	0.1667	19.6%	0.1278	60.4%	25
Base detectors selection among 37 base detectors (experiments of Section V-D)										
$G37a$	0.2411	10.2%	0.1283	33.8%	13.02	0.1506	8%	0.1217	52.7%	13.69
$G37b$	0.2633	20.4%	0.1307	36.3%	**12.74**	0.1659	19%	0.1329	66.8%	**11.78**
$G37$	**0.2743**	**25.4%**	**0.138**	**43.9%**	13	**0.1760**	**26.3%**	**0.1361**	**70.8%**	12
$R37$ [13]	0.2317	5.9%	0.1202	25.3%	37	0.1387	-0.5%	0.0983	23.3%	37
$ABC37$ [40]	0.2635	20.5%	0.1274	32.9%	12.96	0.1639	17.6%	0.1296	62.6%	12.16
$GA37$ [25]	0.2648	21.1%	0.134	39.7%	37	0.1681	20.6%	0.1341	68.3%	37

another variation of the proposed technique in which only the first step of Alg. 2 (i.e. the complete Alg. 1) is executed and the chromosome whose performance is maximum at the end of that step is returned as the chosen configuration, iii) $G25$, which corresponds to the complete approach proposed in Section IV-B, iv) $ABC25$, which uses the Artificial Bee Colony (ABC) [40] algorithm for base detector selection, v) $R25$, in which the 25 available confidence scores are linearly combined using weights determined through regression, as proposed in [13], and vi) $GA25$, in which the 25 available confidence scores are linearly combined using weights determined through gradient ascent, as proposed in [25]. All approaches are executed independently for each concept of both the 2011 and 2012 TRECVID datasets. Subsequently, the above 6 experiments are repeated, this time starting from the 37 base detectors used in $A37$ (thus, 37 base detectors are available for selecting a subset of them, as opposed to 25 in the previous 6 experiments). These experiments are denoted $G37a$, $G37a$, $G37$, $ABC37$, $R37$, $GA37$, respectively. For running the above experiments, the Alg. 1 parameters were set to $N = 5000$, $R = 60$, $k = 40$, $m = 0.15$ and $T = 500$, while the employed Artificial Bee Colony (ABC) parameters were the ones proposed in [40].

The results of the experiments (MXInfAP) are reported in the second and third block of Table III, together with the corresponding average performance boost compared to $A13$ and the mean number of base detectors that were used in each experiment. The latter number is assumed to be approximately proportional to the computational cost of the overall concept detection pipeline at run time.

Firstly, the results show that the $G25$ and $G37$ approaches compare favorably to the $A13$, $A25$ and $A37$ ones examined

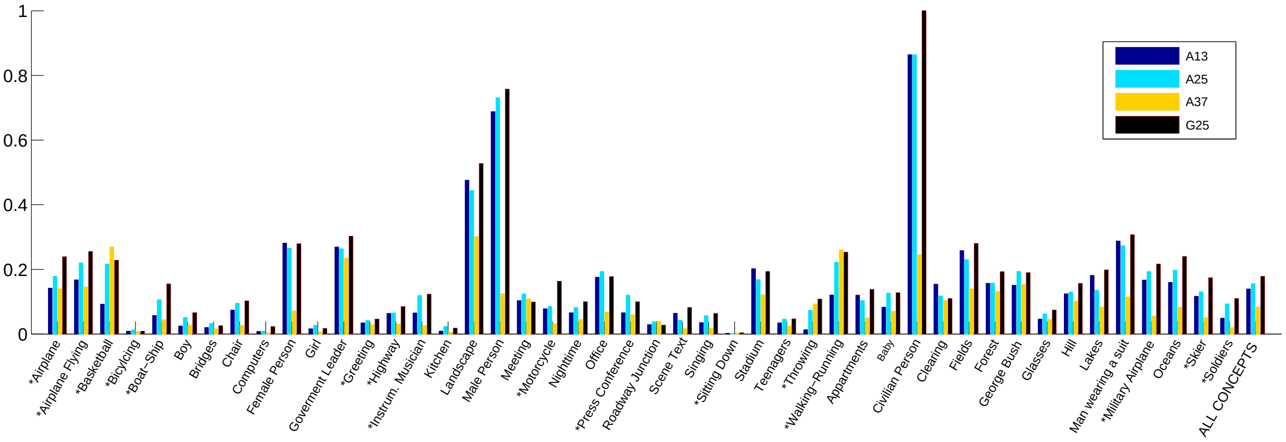
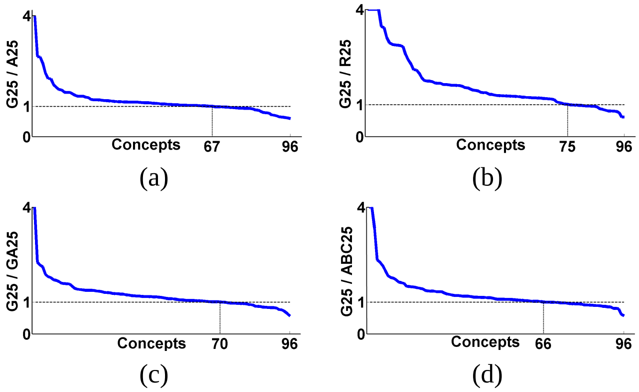


Fig. 5. XIinfAP for the concepts of the 2012 dataset. The compared techniques are $A13$, $A25$, $A37$ and $G25$.

in Section V-C, thus demonstrating the impact of the proposed base detector selection strategy to both MXInfAP and computational cost. Secondly, we can see that the results of $G25$ and $G37$ are very similar (the MXInfAP difference between $G25$ and $G37$ is only 0.0001 and 0.0023 in the 2011 and 2012 dataset). Following the discussion in Section V-C on how Harris-Laplace interest point detection in tomographs leads to noisy base detectors, this similarity of results shows that the proposed base detector selection strategy can effectively identify and discard noisy detectors. Further to this, we can see in Table III that when considering motion-related concepts only, $G37$ outperforms $G25$; this demonstrates the merit of even noisy base detectors under certain conditions. $G25$ is further compared with $A13$, $A25$ and $A37$ for each concept separately, and the results are shown in Figs. 4 and 5, where it can be seen that $G25$ improves on the results of $A13$, $A25$ and $A37$ for the vast majority of the concepts.

Comparison of $G25$ and $G37$ with simpler variants of them, namely $G25a$, $G25b$, $G37a$, $G37b$ (Table III), shows the significance of proposed Algs. 1 and 2 in making a stable selection of good base detectors for each concept. Particularly the importance of taking into account not just the $top-T$ but also the $bottom-T$ configurations when selecting base detectors (second step of Alg. 2) is made clear by this comparison.

With respect to other SoA methods (Table III), the proposed $G25$ (or $G37$) approach compares favorably to $ABC25$ [40] (or $ABC37$), a metaheuristic search algorithm for optimization. This can be explained by the fact that the ABC technique does not take into account that in the concept detection task the solution space is noisy, thus the optimization in a validation set often does not lead to good detection in the testset. Similar conclusions can be drawn by comparing $G25$ (or $G37$) with $GA25$ [25] (or $GA37$) and $R25$ [13] (or $R37$), the most recent base detector combination approaches that can be found in the relevant literature. It seems that the performance reached through gradient ascent (i.e. $GA25$) is similar to the performance reached through a simple genetic algorithm ($G25b$) or ABC ($ABC25$). However, linear combination approaches by definition use the confidence scores of all base detectors. Consequently, both $GA25$ ($GA37$) and $R25$ ($R37$) are much more computationally demanding at runtime than the corresponding algorithms that use a base detector selection scheme (e.g. $G25$). Finally, the poor results of $R25$ ($R37$) can be explained by the fact that for most concepts the number of positive samples in the dataset is much lower than the number of negative samples. This bias undermines this technique, which minimizes the total sum of the error without discriminating between positive and negative samples.



(a)

(b)

(c)

(d)

Fig. 6. Comparison of $G25$ against $A25$, $R25$, $GA25$ and $ABC25$ at the individual concept level, using the ratio of XInfAP achieved by $G25$ and (a) $A25$, (b) $R25$, (c) $GA25$ and (d) $ABC25$. In each sub-figure, the XInfAP ratio is estimated for each of the 96 concepts and then used for sorting the concepts in descending order. Ratios higher than 4 are truncated to 4 for visualization purposes. A dotted horizontal line signifies a ratio equal to 1, while the number of concepts with ratio higher than 1 is highlighted on the $x-axis$.

Further comparison of $G25$ against $A25$, $R25$, $GA25$ and $ABC25$ at the individual concept level (Fig. 6) shows that the former outperforms the latter for 67, 75, 70 and 66 of the 96 concepts, respectively (and is outperformed by only a small margin for most of the few remaining concepts). These results manifest that, although the proposed optimization approach does not theoretically guarantee optimal base detector selection (as does neither one of the related literature approaches), in practice it consistently performs well in these 96 optimization problems.

E. Assessing the impact of algorithm parameters and design choices

A critical parameter of the approach of Section IV is the number of base detectors that are selected for each classifier. We have plotted the MXInfAP of $G25$ and $G37$ as a function of this number (Fig. 7). The plot demonstrates that the achieved MXInfAP does not significantly depend on the exact base detector number, since in all cases there is a rather wide area around its optimal value where it remains almost constant. More specifically, in the $G25$ case the optimal base detector number is 13 in the 2011 dataset and 8 in the 2012 one. The proposed strategy, which is to automatically select the number of base detectors by the genetic algorithm outcome, leads to the selection of 10 and 9 base detectors, respectively. The MXInfAP loss as a result of this is 0.006 in the 2011 dataset (since 0.2804 is the theoretical maximum as seen in Fig. 7, and 0.2744 is the one actually achieved) and 0.0005 in the 2012 dataset (0.1788 and 0.1783 respectively). In the $G37$ case, for both 2011 and 2012 datasets the optimal base detector number is equal to 13, while the proposed strategy leads to the selection of 13 base detectors in the 2011 dataset and 12 base detectors in the 2012 dataset. In the second case, the corresponding MXInfAP loss is 0.0012 (from 0.1772 to 0.176).

For completeness, it should be noted that if a variable number of base detectors is used for each concept, then the overall MXInfAP may further increase. In experiments where we varied the value of M_i between 6 and 25 separately for each concept and selected the optimal configurations, the maximum achieved MXInfAP was 0.2938 and 0.1858 for the 2011 and 2012 dataset, respectively (representing an up to 7% boost over the $G25$ results reported in Table III).

In order to validate our design choice of defining the base detector quality Q_c according to Eq. 4, for selecting with the help of the genetic algorithm the optimal set of base detectors, we further examine the impact of using, in place of Q_c, one of three other reasonable base detector quality measures Q_c^1, Q_c^2, Q_c^3, defined as follows:

$$Q_c^1 = AP_T^P - AP_T^N \tag{5}$$

$$Q_c^2 = P_T/T \tag{6}$$

$$Q_c^3 = AP_T^P \tag{7}$$

where P_T (respectively, N_T) is the number of times that the base detector was included in the configurations that achieved the $top{-}T$ ($bottom{-}T$) performance, and AP_T^P (respectively, AP_T^N) is the Average Precision value that can be calculated by treating the list of $top - T$ ($bottom - T$) configurations as a ranked list of retrieval results, where configurations that include the base detector in question are taken as positive samples. The results, demonstrated in Table IV, manifest the validity of a genetic algorithm that looks at not only the $top - T$ but also the $bottom - T$ configurations, since Q_c and Q_c^1 outperform the corresponding quality measures that employ only the $top - T$ configurations, Q_c^2, Q_c^3. Additionally, averaging the results (Q_c, Q_c^2) seems to exhibit more stable performance than taking into account the exact rank by calculating some sort of Average Precision (Q_c^1, Q_c^3), although in general all observed differences are rather small, suggesting that the proposed genetic algorithm is relatively insensitive to the exact way in which Q_c is defined.

TABLE IV
PERFORMANCE COMPARISON WHEN EACH OF THE FOUR DIFFERENT QUALITY MEASURES DEFINED IN EQS. 4 - 7 ARE USED.

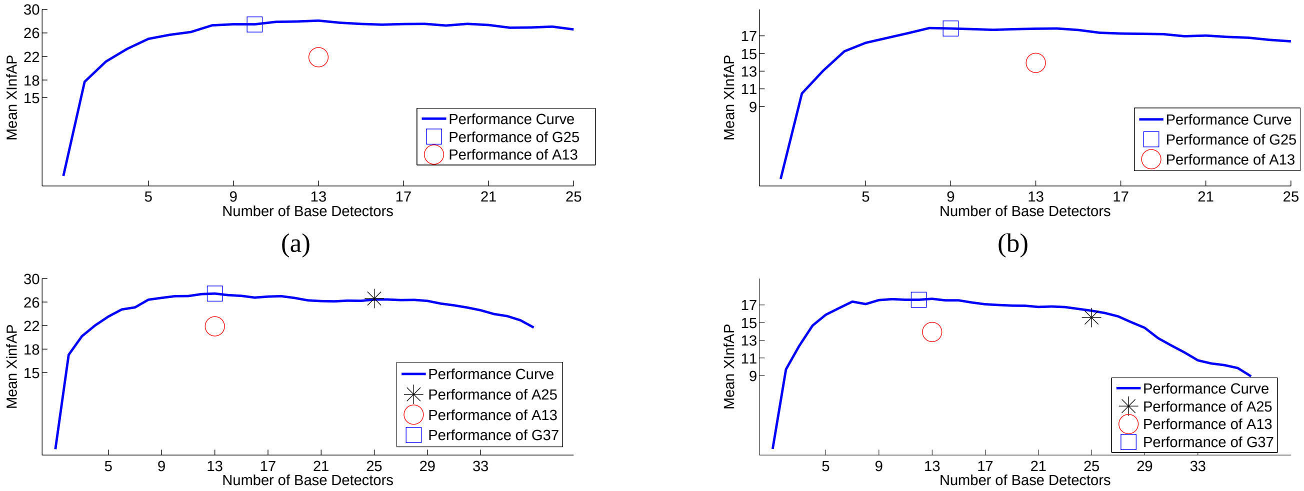
Method	Dataset	Q_c (4)	Q_c^1 (5)	Q_c^2 (6)	Q_c^3 (7)
$G25$	2011	**0.2744**	0.2731	0.2736	0.2728
$G25$	2012	**0.1783**	0.1762	0.1691	0.1692
$G37$	2011	**0.2743**	0.2714	0.2718	0.2705
$G37$	2012	0.1760	**0.1762**	0.1706	0.1680

F. Statistics on the usefulness of different base detectors

Figure 8 shows the number of times that each base detector is used in all 96 concept classifiers if the $G37$ approach is followed. It can be seen that the keyframe-based base detectors are selected more often than the tomograph-based ones. Specifically, on average only 33.8% of the 37 available base detectors are used for a given concept, and 10 out of 13 keyframe-based detectors are consistently selected for more than 50% of the examined concepts. There are also 3 tomograph-based detectors that are selected for the majority of the concepts, despite the fact that motion-related concepts are a minority in our datasets (28 motion-related concepts out of 96 concepts in total). Finally, the tomograph-based base detectors that employ interest point detection via Harris-Laplace rather than extracting local descriptors on a dense grid are the least contributing ones, being ranked in the last 12 places in Fig. 8.

G. Discussion on computational complexity

For comparing the methods' time efficiency throughout the paper, we made the assumption (in Section V-D) that the computational cost of the overall concept detection pipeline at run time is approximately proportional to the mean number of employed base detectors. However, under the $G25$ (and $G37$) approaches, since a different subset of base detectors is used by each concept classifier, features will in any case need to be extracted for all 25 (or 37) possible base detectors when working with multiple concepts. Thus, the above assumption will only hold if the most computationally expensive step of the pipeline is evaluating the trained SVMs, rather than performing feature extraction and BoW creation. Our experiments show that this is indeed the case under the typical application scenario that involves the detection of a plurality of different concepts, e.g. one thousand or more as in [41], [42]. Specifically, the time needed for evaluating the trained SVMs when detecting one thousand concepts under the $A13$ and $G25$ approaches (with 9 base detectors being used for each concept in $G25$) accounts for 97% and 93.9% of the overall processing time, respectively, despite using linear SVMs, as a result of the high number of such SVMs that need to be evaluated. In this experiment, $G25$ is found to be overall 28.6% faster than $A13$,



(a) (b)

(c) (d)

Fig. 7. Performance as a function of the number of selected base detectors. (a) $G25$ in the 2011 dataset (b) $G25$ in the 2012 dataset. (c) $G37$ in the 2011 dataset (d) $G37$ in the 2012 dataset. For better readability, the numbering of the vertical axes shows only the fractional part of the MXInfAP value, i.e. the integer part "0." of MXInfAP is omitted.

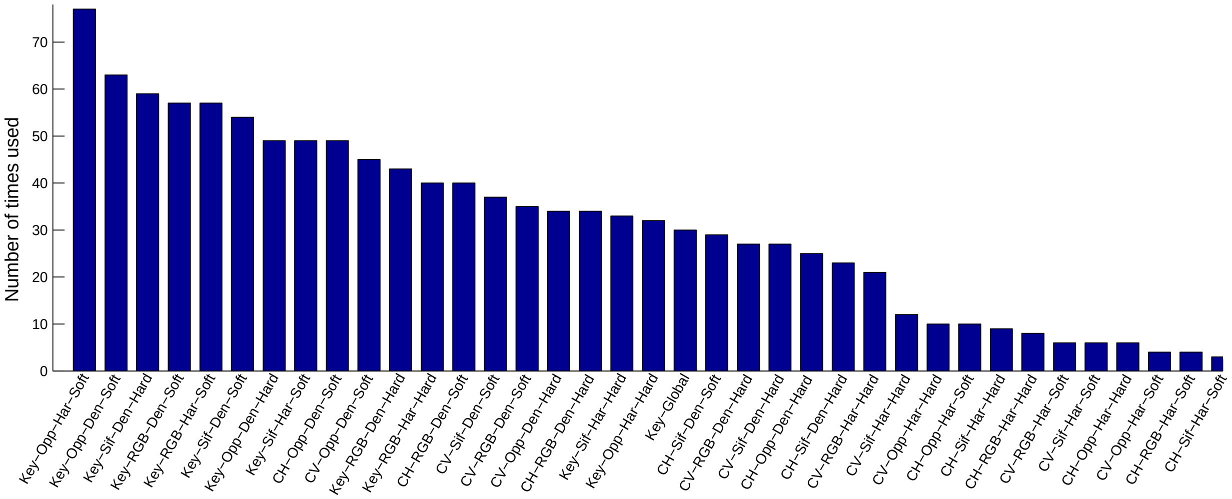


Fig. 8. Barplot of the number of times each base detector is used in all 96 concept classifiers of the two employed datasets, when selection of base detectors is performed according to $G37$. The naming convention for the base detectors (horizontal axis) is <Video sampling>-<Descriptor>-<Interest point detector>-<BoW assignment strategy>. "Key-Global" refers to the HSV histogram-based base detector. See Table II for details on the base detectors.

as a result of using 30.8% fewer base detectors per concept, despite feature extraction and BoW creation in $G25$ requiring somewhat more time than in $A13$.

These run time efficiency gains of base detector selection come at a cost during the training phase, which however is performed off-line and only once for each concept, thus not affecting the scalability of the developed detectors, i.e. their applicability to extremely large volumes of non-labeled video data. Moreover, using the proposed genetic algorithm the training cost is limited because this algorithm visits only a very small portion of all possible configurations. Indicatively, in the most extensive base detector setup ($G37$) the number of possible configurations is 13.7 billion (i.e. 2^{37}), while the configurations that were examined by the introduced algorithm during our $G37$ experiment (in all stages) were approximately 100 thousand. Examining a single configuration using our MATLAB implementation of the base detector selection algorithm on an INTEL(R) Core i7-3770K 3.5GHz PC took 0.7 sec., leading to a total training time of less than 20 hours for each concept; this is certainly not prohibitive for an off-line training process that needs to be performed only once.

VI. Conclusion

In this work we dealt with large-scale video concept detection. We showed that video tomographs can contribute to increased concept detection accuracy, both for motion-related and non-motion-related concepts, while introducing no greater computational cost than that of processing a single keyframe. Additionally, we showed that the "one size fits all" approach for the detection of multiple concepts in video streams leads to the unnecessary evaluation of base detectors that do not always contribute to the detection of a specific concept, thus compromising accuracy while increasing the computational cost. We addressed this issue with a new approach that builds upon a genetic algorithm to rank base detectors and select only a (different for each concept) subset of them. The reported results demonstrate that by combining our base detector selection technique and video tomographs, concept detection effectiveness can be boosted by $25 - 28\%$ (as measured by MXInfAP), while at the same time the run-time computational cost of concept detection can be decreased as a result of using up to $23 - 30\%$ fewer base detectors per concept, compared to indiscriminately using the complete set of keyframe-based base detectors.

REFERENCES

[1] A. Smeaton, P. Over, and W. Kraaij, “High-Level Feature Detection from video in TRECVID: a 5-year retrospective of achievements,” in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[2] U. Niaz, M. Redi, C. Tanase, B. Merialdo, G. Farinella, and Q. Li, “Eurecom at TRECVID 2011: The light semantic indexing task,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, December 2011.

[3] C. Snoek, K. van de Sande, X. Li, M. Mazloom, and et al., “The MediaMill TRECVID 2011 semantic video search engine,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, December 2011.

[4] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] K. van de Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[6] H. Bay, T. Tuytelaars, and L.-V. Gool, “Surf: Speeded up robust features,” in *Proc. European Conf. on Computer Vision*, 2006, pp. 404–417.

[7] E. Tola, V. Lepetit, and P. Fua, “A fast local descriptor for dense matching,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[8] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.

[9] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. Journal of Computer Vision*, vol. 72, no. 2, pp. 213–238, 2007.

[10] Y. G. Jiang, C. W. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” in *Proc. 6th ACM Int. Conf. on Image and video retrieval*, 2007, pp. 494–501.

[11] S. Maji, A. C. Berg, and J. Malik, “Classification using intersection kernel support vector machines is efficient,” in *Proc. IEEE Con. on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[12] N. Inoue and K. Shinoda, “A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors,” *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 1196–1205, 2012.

[13] M. Hradis, I. Reznicek, K. Behun, and L. Otrusina, “Brno university of technology at TRECVID 2011 SIN, CCD,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, December 2011.

[14] Y. Tonomura and A. Akutsu, “Video tomography: An efficient method for camerawork extraction and motion analysis,” in *Proc. ACM Int. Conf. on Multimedia (ACM MM 1994)*, 1994, pp. 349–356.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.-V. Gool, “A comparison of affine region detectors,” *Int. Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[16] I. Laptev and T. Lindeberg, “Space-time interest points,” in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2003, pp. 432–439.

[17] G. Willems, T. Tuytelaars, and L.-V. Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” in *Proc. 10th European Conf. on Computer Vision (ECCV)*, 2008, pp. 650–663.

[18] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[19] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *Proc. British Machine Vision Conference (BMVC)*, 2009.

[20] M.-Y. Chen and A. Hauptmann, “MoSIFT: Recognizing human actions in surveillance videos,” School of Computer Science, Carnegie Mellon University, Tech. Rep., 2009.

[21] H. Jegou, M. Douze, C. Schmid, and P. Perez, “Aggregating local descriptors into a compact image representation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3304–3311.

[22] F. Jurie and B. Triggs, “Creating efficient codebooks for visual recognition,” in *Proc. 10th Int. Conf. on Computer Vision (ICCV)*, 2005, pp. 604–610.

[23] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, “Learning automatic concept detectors from online video,” *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 429–438, 2010.

[24] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui, “Short-term audio-visual atoms for generic video concept classifications,” in *Proc. 17th ACM Int. Conf. on Multimedia*, 2009, pp. 5–14.

[25] S. Sato, N. Inoue, Y. Kamishima, T. Wada, and K. Shinoda, “Tokyotech+canon at trecvid 2011,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, December 2011.

[26] A. Hauptmann and M. Smith, “Text, speech, and vision for video segmentation: The informedia project,” in *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, 1995.

[27] W. Jiang and A. Loui, “Video concept detection by audio-visual grouplets,” *Int. Journal of Multimedia Information Retrieval*, vol. 1, no. 4, pp. 223–238, 2012.

[28] G. Leon, H. Kalva, and B. Furht, “Video identification using video tomography,” in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2009, pp. 1030–1033.

[29] H.-S. Min, S. Kim, W. Neve, and Y. Ro, “Video copy detection using inclined video tomography and bag-of-visual-words,” in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2012, pp. 562–567.

[30] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris, “Enhancing video concept detection with the use of tomographs,” in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 2013.

[31] T. Yu and Y. Zhang, “Retrieval of video clips using global motion information,” *Electronics Letters*, vol. 37, no. 14, pp. 893–895, 2001.

[32] G. Rath and A. Makur, “Iterative least squares and compression based estimations for a four-parameter linear global motion model and global motion compensation,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1075–1099, 1999.

[33] V. Mezaris, I. Kompatsiaris, N. Boulgouris, and M. Strintzis, “Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606–621, 2004.

[34] E. Yilmaz, E. Kanoulas, and J. A. Aslam, “A simple and efficient sampling method for estimating AP and NDCG,” in *Proc. 31st Int. ACM SIGIR Conf. on research and development in information retrieval*, 2008, pp. 603–610.

[35] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, “Large-scale multimodal semantic concept detection for consumer video,” in *Proc. Int. Workshop on Multimedia Information Retrieval (MIR)*, 2007, pp. 255–264.

[36] P. Over, G. Awad, J. Fiscus, B. Antonishek, and et al., “Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, December 2011.

[37] P. Over, J. Fiscus, G. Sanders, B. Shaw, and et al., “Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” in *Proc. TRECVID Workshop*, Gaithersburg, USA, November 2012.

[38] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek, “Visual word ambiguity,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[39] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 2169–2178.

[40] D. Karaboga and B. Basturk, “A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm,” *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, 2007.

[41] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proc. ACM Multimedia*, Barcelona, Spain, October 2013, p. 223.

[42] M. Mazloom, A. Habibian, and C. Snoek, “Querying for video events by semantic signatures from few examples,” in *Proc. ACM Multimedia*, Barcelona, Spain, October 2013, p. 609.